HEC MONTRÉAL

Analyse spatiale appliquée aux inondations du Québec : le cas de la Croix-Rouge canadienne

par

Enora Georgeault

Aurélie Labbe
Marie-Ève Rancourt
HEC Montréal
Directrices de recherche

Sciences de la gestion
(Spécialisation Science des données et analytique d'affaires)

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences
(M. Sc.)

Août 2021

HEC MONTRÉAL

Retrait d’une ou des pages pouvant contenir des renseignements personnels

Résumé

Selon l'Agence de santé publique du Canada, la configuration variable des pluies, les tempêtes plus violentes, la fonte plus rapide des neiges et l'élévation du niveau de la mer en raison des changements climatiques vont accroître le risque d'inondation au Canada. Avec les feux de forêts, aussi en hausse, ce sont les catastrophes naturelles qui provoquent le plus de dégâts dans le pays. Acteur clé pour soutenir les individus et les communautés vulnérables, la Croix-Rouge canadienne (CRC) est au premier plan lorsque des catastrophes touchent la population. Dans ce contexte, ses actions visent à atténuer les répercussions des désastres naturels sur les personnes sinistrées et dépendent des capacités de l'organisation à planifier les opérations logistiques de secours. Devant ces réalités, il devient primordial d'améliorer, en amont, la compréhension des facteurs de vulnérabilité et de besoins des sinistrés afin de maximiser l'efficacité des interventions de secours et d'allocation des dons.

L'objectif de ce projet est d'analyser les données relatives à l'aide financière apportée aux victimes des inondations printanières de la province du Québec en 2019. Plus précisément, cette étude vise à détecter la présence de facteurs ou de phénomènes ayant un impact significatif sur la demande d'aide des foyers sinistrés et par conséquent sur l'allocation des dons de la CRC. Pour ce faire, nous avons étudié, à l'aide de différentes méthodes d'analyse spatiale, les données récoltées par la CRC couplées à des données externes pouvant expliquer les allocations financières. Différents angles d'étude ont été abordés selon le choix du niveau d'agrégation des données. Ces analyses suggèrent un lien significatif entre le nombre de personnes dans un foyer et son montant reçu. De plus,

l'analyse surfacique a mis en lumière la présence d'interactions endogènes et exogènes lors de la modélisation du montant moyen reçu par aire de diffusion. En d'autres termes, les valeurs des régions voisines s'influencent entre-elles et le montant moyen versé d'une zone est donc dépendant de ses propres caractéristiques mais aussi de celles de ses voisines.

Cependant, qu'il s'agisse des résultats de l'analyse des données ponctuelles (sans agrégation) ou de ceux relatifs à l'analyse surfacique (agrégation par aire de diffusion), les résultats sont mitigés. Tout d'abord, si la faible variabilité dans les montants versés aux sinistrés s'explique par le versement d'aides à grande échelle basées sur un principe de neutralité, elle constitue cependant un obstacle de taille pour répondre à la question de recherche. Ensuite, la difficulté liée à la collecte de données de qualité en situation d'urgence est un enjeu majeur pour la Croix-Rouge et les organisations humanitaires. Ces aspects compliquent l'obtention de résultats significatifs et impliquent une prudence nécessaire quant à l'interprétation des résultats.

Mots-clés

Analyse spatiale, données ponctuelles, données surfaciques, processus ponctuels, autocorrélation, modèles d'économétrie spatiale, inondations

Table des matières

Résumé **iii**

Liste des tableaux **ix**

Liste des figures **xi**

Liste des abréviations **xv**

Remerciements **xvii**

Introduction **1**

1 Introduction à l'analyse des données spatiales **3**

1.1 Les trois types de données spatiales 3

1.1.1 Les données surfaciques 4

1.1.2 Les données ponctuelles 4

1.1.3 Les données continues ou géostatistiques 5

1.2 Notions et spécificités des données spatiales 6

1.2.1 Le Problème des Unités Spatiales Modifiables - MAUP 7

1.2.2 Fenêtre d'observation et traitement des effets de bord 7

1.2.3 Systèmes de coordonnées de référence 8

1.3 Les différents outils disponibles 9

2 Analyse des données ponctuelles **11**

2.1 Introduction au concept des processus ponctuels 12
2.1.1 Processus de Poisson homogène *(Complete Spatial Randomness, CSR)* 13
2.1.2 Processus de Poisson non-homogène 13
2.2 Analyses exploratoires 14
2.2.1 Intensité 14
2.2.2 Dépendance et corrélation 22
2.3 Processus marqués 29
2.3.1 Analyse exploratoire 29
2.4 Régression géographiquement pondérée (RGP) 34
2.4.1 Estimation et paramètres du modèle 36

3 Analyse des données surfaciques 41
3.1 Définition d'une structure de voisinage 42
3.1.1 Définition basée sur la contiguïté 43
3.1.2 Définition basée sur la distance 43
3.1.3 Autres méthodes 46
3.1.4 Attribution de poids 47
3.2 Autocorrélation spatiale 49
3.2.1 Définition de l'autocorrélation spatiale 49
3.2.2 Dépendance spatiale globale 50
3.2.3 Dépendance spatiale locale 54
3.3 Modélisation 56
3.3.1 Modèles d'économétrie spatiale 56
3.3.2 Tests et critères statistiques pour le choix du modèle 58
3.3.3 Interprétation des résultats 61

4 Application aux données d'inondation de la Croix-Rouge canadienne 63
4.1 Présentation des données et étapes préliminaires 64
4.1.1 Données de la Croix-Rouge canadienne (CRC) 64

4.1.2 Données externes 66
4.1.3 Pré-traitement pour joindre les données externes aux données de la CRC 68
4.1.4 Création de sous-ensembles 71
4.1.5 Analyse descriptive des variables 72
4.2 Analyse des données ponctuelles 76
4.2.1 Estimation de l'intensité par la méthode des quadrats 76
4.2.2 Étude de la dépendance avec la fonction K de Ripley 81
4.2.3 Analyse des marques 82
4.2.4 Régression géographiquement pondérée 86
4.3 Analyse des données surfaciques 93
4.3.1 Description des données 93
4.3.2 Définition de la structure de voisinage 94
4.3.3 Autocorrélations globale et locale 96
4.3.4 Modélisation 101

Conclusion **109**

Bibliographie **115**

Annexe A – Analyse descriptive des variables **i**

Annexe B – Régression géographiquement pondérée **v**

Annexe C - Analyse surfacique **vii**

Liste des tableaux

1.1 Liste non-exhaustive des principaux paquetages pour l'analyse des données spatiales en R. 10

2.1 Liste des jeux de données utilisés dans le chapitre 2. 12
2.2 Fonctions noyau. 37
2.3 Descriptions des variables du jeu de données *EWHP* présent dans le paquetage *GWmodel*. 39
2.4 Résultats de la régression linéaire et de la régression géographiquement pondérée. 40

4.1 Liste des variables utilisées provenant du Recensement de 2016 par Statistique Canada. 68
4.2 Noms des subdivisions de recensement utilisées pour sélectionner les foyers pour les sous-ensembles de points *Montreal* et *Marthe*. 72
4.3 Statistiques descriptives des variables disponibles pour le jeu de données complet (Québec) de 4705 observations après pré-traitement. 77
4.4 Bandes passantes optimales selon différentes méthodes disponibles dans *spatstat*. 79
4.5 Résultats des modèles de régression linéaire classique selon les variables utilisées. 88
4.6 Performances des modèles selon les valeurs des bandes passantes optimisées pour le jeu de données du Québec. 90

4.7 Statistiques descriptives des coefficients des RGP (noyau exponentiel fixe) avec les performances des modèles associés. 91
4.8 Statistiques descriptives pour les 177 aires de diffusion ayant au moins cinq observations. 96
4.9 Analyse descriptive des structures de voisinage pour les 177 aires de diffusion (AD) liées à au minimum cinq foyers sinistrés. 97
4.10 Statistiques descriptives des indices de Moran locaux et de leur valeurs-p. . . 100
4.11 Distribution des valeurs-p des indices de Moran locaux, selon la méthode d'ajustement utilisée. 101
4.12 Résultats du test de Moran adapté sur les résidus du modèle RLC selon la matrice de voisinage (normalisation en ligne). 104
4.13 Déterminants du montant moyen alloué à chaque foyer par aire de diffusion, à partir d'une matrice basée sur la notion contiguïté (Queen) avec normalisation en ligne (W). 105
4.14 Impacts directs, indirects et totaux du modèle SDM (matrice de contiguïté avec normalisation en ligne). 108

1 Statistiques descriptives des variables disponibles pour le jeu de données *Montréal* de 2406 observations après pré-traitement. ii
2 Statistiques descriptives des variables disponibles pour le jeu de données *Marthe* de 1222 observations après pré-traitement. iii
3 Fréquence des variables représentant les tranches d'âge sur le jeu de données total (Québec) de 4705 observations après pré-traitement. iii

4 Performances des modèles de régression linéaire et de RGP selon les valeurs des bandes passantes optimisées pour le jeu de données *Montréal*. vi

5 Indices de Moran (Global) selon la définition du voisinage et le type de normalisation. viii

Liste des figures

1.1 Localisation des nids de fourmis selon deux espèces et localisation et diamètre d'anémones de mer. 6

2.1 Processus de Poisson homogène : simulation de 4 configurations de points avec une intensité de 10 points par unité de surface. 14
2.2 Dénombrements par la méthode des quadrats et mesure d'intensité pour le jeu de données *swedishpines* 17
2.3 Dénombrements par la méthode des quadrats hexagonaux et mesure d'intensité pour le jeu de données *swedishpines* 18
2.4 Estimation de la densité du jeu de donnée *swedishines* selon différentes valeurs de bande passante pour un noyau Gaussien avec une correction « uniforme ». 22
2.5 Représentation des trois configurations-types : agrégée, régulière et aléatoire. 24
2.6 Les fonctions K des configurations-types (présentées à la figure 2.5) 26
2.7 Configuration de points avec une intensité non homogène (gauche), la courbe de la fonction K associée (milieu) et la courbe de la fonction K_{inhom} (à droite) 27
2.8 Localisations et diamètres d'anémones de mer. 30
2.9 Emplacement de 126 arbres (jeunes pousses de pins) dans une forêt finlandaise et nuage de points de leurs diamètres. 31
2.10 Lissage des marques du jeu de données 32
2.11 Estimation de la fonction de corrélation de marque (diamètre) pour le jeu de données *anemones* disponible dans *spatstat*. 34

2.12 Estimation de la fonction *K* pondérée par les marques (diamètre) pour le jeu de données *anemones* disponible dans *spatstat*. 35
2.13 Visualisation des estimations des coefficients de la variable *FlrArea* de la RPG (avec noyau adaptatif Bicarré et bande passante optimisée selon le AIC) du jeu de données *EWHP* présent dans le paquetage *GWmodel*. 40

3.1 Visualisation du nombre de cas de leucémie par secteur de recensement pour la ville de Syracuse de l'état de New-York, États-Unis. 42
3.2 Contiguïté QUEEN et ROOK pour la ville de Syracuse de l'état de New-York, États-Unis. 44
3.3 Graphes de voisinage fondés sur des notions géométriques pour la ville de Syracuse de l'état de New-York, États-Unis. 45
3.4 Graphes de voisinage fondés sur les plus proches voisins pour la ville de Syracuse de l'état de New-York, États-Unis. 46
3.5 Diagramme de Moran pour la distribution des cas de leucémie dans l'état de New-York, États-Unis. 52

4.1 Visualisation des positions d'inondations significatives selon *Données Québec*. 69
4.2 Visualisation des foyers ayant reçu une aide de la CRC après le traitement des anomalies. 71
4.3 Distribution selon les catégories d'âge et de sexe du nombre de personnes dans les foyers pour le jeu de donnée complet *Québec*. 74
4.4 Distribution du montant d'argent reçu par foyer selon le niveau de sévérité. Les valeurs extrêmes n'ont pas été prises en compte pour ce diagramme. . . . 75
4.5 Corrélations entre le montant reçu par foyer (*Montant reçu*) et les autres variables disponibles. 76
4.6 Analyse de l'intensité par la méthode des quadrats (zone de Montréal). 78
4.7 Analyse de l'intensité par la méthode des quadrats (zone de Sainte-Marthe-sur-le-Lac). 79
4.8 Estimation de la densité de la zone *Marthe* selon plusieurs noyaux. 80

4.9 Représentation des Fonctions K_{inhom} (courbes noires) pour les trois zones étudiées (Sainte-Marthe-sur-le-Lac, Grand Montréal et la province du Québec). . 81

4.10 Répartition spatiale des montants reçus en $ CAD pour le jeu de données *Québec*. 83

4.11 Répartition spatiale des montants reçus en $ CAD pour le jeu de données *Montréal*. 83

4.12 Résultat du lissage de l'intensité de la marque du montant reçu en $ CAD pour *Québec* selon la taille des bandes passantes. 84

4.13 Résultat du lissage de l'intensité de la marque du montant reçu en $ CAD pour *Montréal* selon la taille des bandes passantes. 84

4.14 Résultat du lissage de l'intensité de la marque du montant reçu en $ CAD pour *Marthe* selon la taille des bandes passantes. 85

4.15 Fonction de corrélation de marque (montant reçu) et fonction *K* pondérée par la marque (montant reçu) pour le jeu de données du *Québec*. 86

4.16 Fonction de corrélation de marque (montant reçu) et fonction *K* pondérée par la marque (montant reçu) pour le jeu de données du *Montréal* 87

4.17 Fonction de corrélation de marque (montant reçu) et fonction *K* pondérée par la marque (montant reçu) pour le jeu de données du *Marthe* 87

4.18 Répartition géographique et distribution des valeurs du coefficient de RGP pour le nombre de personne dans le foyer sinistré. 92

4.19 Estimation continue des valeurs du coefficient de RGP pour le nombre de personnes dans le foyer sinistré. 92

4.20 Visualisation des aires de diffusion associées aux foyers sinistrés (Québec). . 94

4.21 Visualisation des aires de diffusion sélectionnées pour l'analyse des données surfaciques. 95

4.22 Représentation des relations entre aires de diffusion selon la définition de la structure de voisinage. 98

4.23 Diagrammes de Moran des montant moyens par aire de diffusion pour la distribution réelle et pour une distribution simulée aléatoirement avec permutations. 99
4.24 Significativité et structure dominante selon l'indice de Moran local. 102
4.25 Corrélations entre les 10 variables explicatives sélectionnées et la variable d'intérêt *Montant moyen*. 103
4.26 Tests d'hypothèses pour comparer les modèles de régression spatiaux selon la méthode ascendante. 106
4.27 Tests d'hypothèses pour comparer les modèles de régression spatiaux selon la méthode descendante. 106
4.28 Tests d'hypothèses pour comparer les modèles de régression spatiaux selon la méthode mixte. 107

Liste des abréviations

AD Aire(s) de diffusion

AIC Critère d'information d'Akaike (*Akaike Information Criterion*)

CRC Croix-Rouge canadienne

CSR Processus de Poisson homogène (*Complete Spatial Randomness*)

DQ Données Québec

GPS Géo-positionnement par satellite (*Global Positioning System*)

LISA Indicateurs d'autocorrélation spatiale locale

MCO Moindre carrés ordinaire

RGP Régression(s) géographiquement pondérée(s)

RLC Régression(s) linéaire(s) classique(s)

SAR Modèle autorégressif spatial (*Spatial AutoRegression*)

SDM Modèle spatial de Durbin (*Spatial Durbin model*)

SEM Modèle à erreur autocorrélées spatialement (*Spatial Error model*)

SIG Système(s) d'information géographique

SLX Modèle à interactions exogènes (*Spatial Lag X*)

Remerciements

Après plusieurs mois de travail, je souhaite témoigner ma gratitude à de nombreuses personnes indispensables à l'achèvement de ce mémoire.

Tout d'abord, je souhaite sincèrement remercier mes directrices de recherche Aurélie Labbe et Marie-Ève Rancourt de m'avoir fait confiance en me proposant ce passionnant projet de recherche. Je suis très reconnaissante d'avoir pu travailler à leur côté car leur implication, leurs conseils et leurs encouragements ont été très précieux tout au long de ce travail.

Je tiens aussi à remercier les membres de la Croix-Rouge canadienne qui ont permis la réalisation de ce projet. Merci à Madeleine et Peter d'avoir été mes référents et de m'avoir orientée lorsque j'avais des interrogations. Je remercie également toutes les personnes ayant pris le temps de m'informer et de m'éclairer sur leur rôle et, plus généralement, sur le fonctionnement de la CRC.

Je suis aussi profondément reconnaissante d'avoir reçu une bourse de maîtrise de la part d'IVADO. Outre le soutien financier, j'ai aussi eu la chance, grâce à cette bourse, de participer à des formations et des événements très enrichissants organisés par IVADO.

Merci à Geneviève Benoit et à la direction du programme de m'avoir fait confiance et permis de réaliser ce travail tout en étudiant durant ma deuxième année de maîtrise, dans le cadre d'un double-diplôme, à l'ESCP Europe.

Je remercie tous mes amis qui m'ont apporté motivation, soutien et encouragements durant ce projet. Plus particulièrement, merci à Caroline qui, il y a six ans maintenant, m'a motivée et convaincue d'entamer mon parcours universitaire à HEC Montréal, loin

de ma France natale.

Pour finir, j'aimerais particulièrement remercier mes parents et ma soeur qui, depuis toujours, m'accompagnent et me soutiennent de manière inconditionnelle.

Introduction

C'est au Dr John Snow, avec son étude sur la répartition spatiale des cas de choléra à Londres en 1854, que l'on attribue la première contribution à l'analyse des données spatiales. Depuis, les techniques des systèmes d'information géographique (SIG ou *GIS*) ont largement évolué, notamment grâce à l'augmentation du volume de données spatiales disponibles, rendue possible avec le développement des outils de géo-positionnement par satellite (*Global Positioning System*, GPS) dans les années 1970. On a depuis assisté à l'essor des statistiques spatiales, fortement utilisées dans de multiples domaines tels que l'épidémiologie, l'économétrie, les sciences de l'environnement ou encore la prospection minière. Aujourd'hui, avec la géolocalisation, la quantité de données géospatiales récoltées est colossale. Leur valorisation devient alors un enjeu majeur pour de plus en plus de secteurs.

De manière plus globale, l'utilisation des méthodes d'analyses de données (pas uniquement spatiales) se généralise et s'applique maintenant à une multitude de domaines, notamment dans l'humanitaire. On observe de plus en plus d'études liées à l'analyse et à la valorisation des données en contexte de crise humanitaire. Par exemple, avec l'essor de l'utilisation des réseaux sociaux, les données issues de ces plateformes sont devenues des sources cruciales d'informations pour l'organisation et le déploiement des opérations de secours. En effet, Cheong et Cheong (2011), qui ont analysé les données extraites de Twitter dans le cadre des inondations de 2010-2011 en Australie, expliquent que les utilisateurs fournissent des observations quasiment en temps réel sur les scènes de catastrophes (telles que des photos et des images aériennes). On peut aussi citer Dontas et collab. (2017) qui

présentent des solutions afin d'analyser les données issues des réseaux sociaux pour améliorer la prise de décision dans le contexte de la crise des réfugiés Syriens. Il devient alors nécessaire pour les associations humanitaires d'exploiter les données externes disponibles, mais aussi celles qu'elles récoltent, à l'interne, durant les opérations d'urgence et d'allocation de dons.

C'est dans ce cadre de valorisation des données spatiales que s'inscrit le présent projet de recherche, en partenariat avec l'organisme humanitaire de la Croix-Rouge canadienne (CRC). Depuis de nombreuses années, la CRC est un acteur clé pour soutenir les individus et les communautés vulnérables, notamment suite à des catastrophes naturelles, comme des inondations ou des incendies. Elle a notamment joué un rôle essentiel lors des inondations subies par les provinces du Québec, de l'Ontario et du Nouveau-Brunswick en 2019. L'objectif de cette étude est d'analyser les données relatives à l'aide financière apportée aux victimes suite à ces inondations dans la province du Québec. Plus précisément, cette étude vise à détecter la présence de facteurs ou de phénomènes ayant un impact significatif sur la demande d'aide des foyers sinistrés et par conséquent sur l'allocation des dons de la CRC. En effet, une bonne compréhension de ces aspects permettrait de faciliter la planification des opérations logistiques d'aide d'urgence et les appels de financement effectués par la CRC.

Pour mener à bien cette étude, la première étape consiste à présenter les notions fondamentales et les méthodes d'analyses spatiales applicables aux données récoltées par la CRC. Le premier chapitre introduira les différents types de données spatiales ainsi que les éléments essentiels à prendre en compte lors de leur analyse. Ensuite, les chapitres deux et trois traiteront de différentes méthodes d'exploration et de modélisation pouvant s'appliquer respectivement aux données ponctuelles et surfaciques. Pour finir, à l'aide des données de la CRC et de données externes pertinentes à l'égard du contexte d'étude, nous effectuerons des analyses exploratoires puis nous tenterons de modéliser l'aide financière apportée par la Croix-Rouge canadienne aux victimes des inondations de 2019 au Québec.

Chapitre 1

Introduction à l'analyse des données spatiales

L'objectif de ce chapitre est d'introduire les concepts fondamentaux liés aux données spatiales, avant d'étudier leurs méthodes d'analyse aux chapitres 2 et 3. Une description des trois types de données géolocalisées proposés par Cressie (1993) sera suivie du développement de certaines notions inhérentes à l'analyse de ces données. Pour finir, on présentera quelques outils et logiciels permettant d'étudier les données spatiales, avec notamment une liste de paquetages utiles en R.

1.1 Les trois types de données spatiales

Une donnée spatiale est une observation à laquelle est rattachée une information annexe sur sa localisation. On est ainsi capable de situer cette observation dans un espace géographique et sur une carte. Ce renseignement constitue une information potentiellement riche pour l'analyse et grâce à ses coordonnées, il est référencé dans un système d'information géographique. Il existe différents types de données spatiales qui ont été classés par Cressie (1993) selon trois catégories : les données surfaciques, les données ponctuelles et les données continues.

1.1.1 Les données surfaciques

La localisation des observations est considérée comme **fixe** pour les données surfaciques (parfois appelées laticielles), et ce sont les valeurs associées qui seront variables. Les observations sont le plus souvent regroupées selon un ensemble de zones, mais elles peuvent également être des points fixes d'un territoire (mairies, écoles, etc.). Ainsi, si ce type de données est souvent représenté sur des surfaces (régions, pays, arrondissements, etc.), il faut tout de même préciser que ce n'est pas toujours le cas et le terme « surfacique » peut donc être trompeur. Par exemple, il peut s'agir du nombre de mariages par mairie ou bien du nombre d'enfants par écoles. En présence de données surfaciques, on s'interroge sur les relations entre les valeurs des observations voisines. On cherche à évaluer l'influence qu'exercent les observations sur leurs voisines, et à quel point celle-ci est significative. Le chapitre 3 se consacre à l'analyse des données surfaciques, et proposera des méthodes d'exploration et de modélisation que l'on pourra ensuite appliquer aux données de la CRC au chapitre 4.

1.1.2 Les données ponctuelles

Les données ponctuelles se caractérisent par la distribution des observations dans l'espace. On s'intéresse à la localisation géographique de l'observation, et non pas à une quelconque valeur (comme pour les données surfaciques ou continues). Comprendre la disposition spatiale des points est alors l'objectif principal de l'analyse. Une configuration de points est un jeu de données ponctuelles spécifiant les emplacements de chacune des observations liées à un événement ou à une entité. Il s'agit, par exemple, de l'emplacement des arbres dans une forêt, des positions exactes d'accidents de la route dans une région, des lieux précis où des homicides ont été commis ou encore de la localisation des pharmacies au sein d'une ville. L'analyse des données ponctuelles est aussi très utilisée en épidémiologie où on peut par exemple étudier la répartition dans l'espace des personnes atteintes d'une certaine maladie.

Les processus marqués et multitypes

Il est possible qu'aux données ponctuelles soient associées une ou plusieurs caractéristiques, qu'on appelle « marques » d'un point. On parlera alors de processus marqués. En effet, lors de la collecte des données, il est probable qu'une étude ne récolte pas uniquement la localisation d'une observation mais qu'elle enregistre aussi d'autres attributs. Ainsi, en plus d'avoir la localisation exacte des arbres dans une forêt, nous pouvons aussi récolter des données concernant son espèce, sa hauteur, son diamètre, etc. Dans le cadre de notre sujet de recherche, en plus de la position des foyers sinistrés, on dispose par exemple du montant reçu ou encore du nombre de personnes associées à ce foyer.

Les marques peuvent être de nature qualitative ou quantitative. La figure 1.1 présente ces deux possibilités : une configuration de points avec deux espèces de fourmis (gauche) et une configuration spécifiant le diamètre d'anémones de mer (droite). Lorsque la marque est catégorielle on parle aussi de processus multitypes ou de processus multivarié (c'est le cas pour la figure 1.1 de gauche). Ces marques permettent d'amener de nouvelles questions de recherches et de faire des analyses plus poussées. On peut par exemple s'intéresser à la dépendance des espèces (y a-t-il des phénomènes d'attraction/de répulsion entre elles ?) ou bien à leur répartition dans l'espace (leurs intensités sont-elles toutes homogènes ou bien y a-t-il de la ségrégation ?). Si une indépendance entre des points du même type signifie qu'il s'agit d'un processus de Poisson, une indépendance entre des points de types différents n'aura pas la même signification. En effet, une indépendance entre des points de catégorie A et des points de catégorie B signifie que ce sont deux processus ponctuels indépendants, mais cela n'indique rien d'autre sur leur répartition spatiale.

1.1.3 Les données continues ou géostatistiques

On parle de données géostatistiques ou données continues lorsque la valeur du phénomène d'intérêt est répartie de manière continue dans l'espace étudié. En d'autres termes, pour chaque point de la zone étudiée, il existe une valeur pour cette variable. Il peut s'agir par exemple de la température, de la qualité de l'air ou encore de la composition chimique

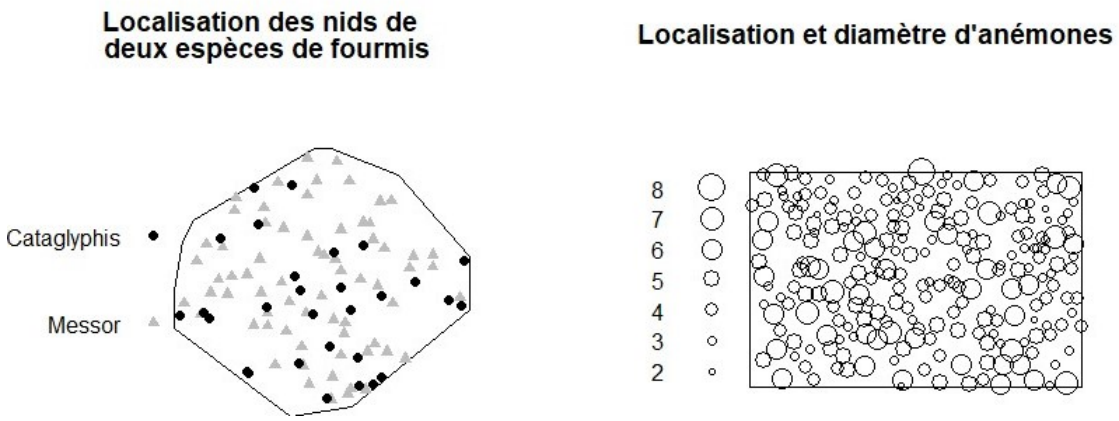


FIGURE 1.1 – Gauche : Localisation des nids de fourmis selon deux espèces : *Cataglyphis* et *Messor*. Jeu de données *ants* disponible dans le package *spatstat*. Droite : Localisation et diamètre d'anémones de mer. Jeu de données *anemones* de *spatstat*.
Sources : Harkness et Isham (1983) (gauche) et Kooijman (1979) (droite).

du sol. Les valeurs de ces variables sont cependant mesurées uniquement en un nombre discret de points. Ainsi, si la distribution spatiale de la température est bien continue, ses valeurs sont néanmoins mesurées grâce à des stations météo situées à des endroits précis du territoire. L'objectif principal de l'étude des données géostatistiques est d'estimer la valeur de la variable d'intérêt dans des endroits où elle n'a pas été échantillonnée. Les applications géologiques et minières sont les plus courantes et sont à l'origine du développement de nombreux concepts en géostatistique, avec par exemple les travaux précurseurs de Matheron (1965). Plus récemment, on peut aussi retrouver le développement de ces notions chez Chiles et Delfiner (2009).

Dans la mesure où ces données ne correspondent pas aux données disponibles dans le cadre de l'étude de cas avec la CRC, nous ne traiterons pas ici des différentes méthodes liées à l'analyse des données continues.

1.2 Notions et spécificités des données spatiales

Cette section a pour but d'introduire différentes notions importantes à prendre en compte lors de l'analyse de données spatiales.

1.2.1 Le Problème des Unités Spatiales Modifiables - MAUP

Le Problème des Unités Spatiales Modifiables (*Modifiable Area Units Problem - MAUP*) est une question importante liée au découpage spatial lorsque l'on analyse des données à un niveau d'agrégation supérieur au niveau d'incidence. Pour Oppenshaw et Taylor (1979), le MAUP correspond à la superposition de l'effet de zonage et de l'effet d'échelle.

- **Effet d'échelle** : Le problème de l'échelle se caractérise par des résultats différents en fonction du niveau d'agrégation des données. Ainsi, plus la taille des unités spatiales est grande, plus les spécificités locales disparaissent au profit de tendances plus globales. Par exemple, les résultats d'une analyse basée sur des données au niveau d'agrégation des provinces peuvent différer de ceux au niveau d'agrégation des secteurs de recensement.

- **Effet de zonage** : Dans la mesure où il existe de multiples façons de découper une région en plusieurs subdivisions, la forme (morphologie et position dans l'espace) des unités spatiales peut aussi avoir une incidence sur les résultats. On peut par exemple mentionner les différents découpages administratifs et électoraux qui représentent bien cette problématique.

Le MAUP est donc une forme d'erreur écologique associée à l'agrégation de données en unités de surface, dont les effets peuvent influencer, par exemple, l'estimation des modèles ou les corrélations. Par conséquent, lorsqu'on traite des données agrégées, il est important de garder à l'esprit que les résultats peuvent être dépendants des unités spatiales utilisées et influencés par leurs configurations et leurs tailles.

1.2.2 Fenêtre d'observation et traitement des effets de bord

Dans le cadre des données ponctuelles et continues, le choix de la fenêtre d'observation, qui désigne les contours de la zone d'étude, est souvent arbitraire. Il peut s'agir par exemple d'une aire d'étude carrée, rectangulaire, circulaire ou encore administrative. La sélection de cette fenêtre d'observation peut ou ne pas être liée au processus étudié.

Par exemple, pour des questions de coûts ou d'opportunités (liés à la collecte des données), on peut avoir à se limiter à une zone d'étude restreinte. À l'inverse, il se peut que les frontières de la fenêtre d'observation soient intrinsèquement liées au phénomène étudié, avec par exemple des limites naturelles (montagnes, fleuves, etc.) qui restreignent géographiquement le processus générateur des données. Dans le premier cas, il est alors nécessaire de prendre en compte cet aspect et on peut par exemple faire l'hypothèse d'une continuité de l'intensité du processus au delà des frontières de la fenêtre. Si au contraire, une frontière naturelle contraint géographiquement le processus observé, alors l'intensité du phénomène observé est nulle à l'extérieur et le traitement des effets de bord n'est pas pertinent. Il existe différentes méthodes pour traiter les effets de bord, ayant tous pour objectif de prendre en compte l'impact de la frontière. Nous les aborderons plus en détails dans les prochaines sections.

1.2.3 Systèmes de coordonnées de référence

Implicitement, toute donnée spatiale est liée à un système de référence spatial permettant de la repérer sur une surface. Il peut s'agir d'un système arbitraire comme par exemple une grille de 10 mètres par 10 mètres dans un champs, ou bien d'un système de référence géographique terrestre.

On peut distinguer deux types de coordonnées qui permettent de se positionner par rapport au système terrestre. En premier, les coordonnées géographiques utilisent la longitude (direction Est-Ouest) et la latitude (direction Nord-Sud) pour situer chaque point sur la surface du globe. La latitude se mesure selon l'angle par rapport au méridien, alors que la longitude est définie par l'angle avec l'équateur. Ainsi, dans le système de coordonnées géographiques, les emplacements sont caractérisés par des degrés et non par des mètres. En second, les systèmes de coordonnées projetés s'appuient sur une modélisation de la terre projetée sur un plan et sur un système de coordonnées cartésien avec un point d'origine, un axe x, un axe y et une unité de mesure.

Lorsque l'on manipule des données spatiales, et notamment lorsque l'on combine

plusieurs types et sources de données, il est nécessaire d'utiliser les mêmes systèmes de coordonnées. Les conversions sont donc indispensables, notamment pour pouvoir superposer des données n'ayant pas les mêmes systèmes de coordonnées. Dans R, le mode *view* du paquetage *tmap* permet de visualiser les données sur un fond de carte interactif (tel que *OpenStreetMap)*, ce qui facilite toute vérification suite à une fusion de fichiers spatiaux avec changement de coordonnées.

1.3 Les différents outils disponibles

Il existe différents outils disponibles pour l'analyses des données spatiales. Parmi les logiciels de système d'information géographique (SIG) les plus connus, on retrouve ArcGIS (sous licence) et QGIS (source ouverte) qui comportent de nombreux outils pour stocker, analyser et cartographier des données spatiales et géographiques. On peut également citer GeoDa qui est aussi un logiciel de source ouverte permettant d'analyser et de visualiser les données géolocalisées.

Ensuite, il est aussi possible d'utiliser des outils flexibles basés sur des langages statistiques et informatiques comme R ou Python (avec notamment les paquetqges *Geopandas, SciPy, PySAL*), qui eux, ne sont pas uniquement orientés sur le traitement des données spatiales. Concernant R, il s'agit sans doute du logiciel le plus complet pour les statistiques spatiales, notamment pour l'estimation des modèles d'économétrie spatiale (Floch et Le Saout, 2018). Dans la mesure où toutes les analyses présentées dans cette études ont été effectuées en R, le tableau 1.1 présente quelques paquetages R incontournables en analyse spatiale.

Paquetages R	**Descriptions et rôles principaux**
`sp`	Classes d'objets spatiaux ; manipulation et visualisation des données
`sf`	Classes d'objets spatiaux ; manipulation et visualisation des données ; version de `sp` plus rapide et flexible
`maptools`	Importation de fichiers shapefile ; manipulation des données spatiales
`spdep`	Création de structures de voisinage, calcul de tests statistiques
`spatstat`	Analyse des données ponctuelles
`dbmss`	Analyse des données ponctuelles
`gstat`	Analyse des données géostatistiques
`rgdal`	Importation de fichiers, manipulations des formats, transformation et manipulation des projections géographiques
`GWmodel`	Modèles géographiquement pondérés
`tmap, ggmap, leaflet`	Visualisation et représentation cartographique

TABLEAU 1.1 – Liste non-exhaustive des principaux paquetages pour l'analyse des données spatiales en R.

Chapitre 2

Analyse des données ponctuelles

L'objectif de ce chapitre est d'introduire les notions fondamentales liées à l'analyse des données ponctuelles et de présenter des méthodes pouvant par la suite être appliquées aux données d'inondations récoltées par la CRC (chapitre 4). Ce chapitre s'appuie en partie sur des exemples et concepts proposés par les paquetages R *spatstat* (Baddeley et collab., 2015) et *GWmodel* (Gollini et collab., 2015), ainsi que sur le chapitre dédié du manuel d'analyse spatiale de l'Institut National de la Statistique et des Études Économiques (Floch et collab., 2018). Après avoir proposé une introduction aux concepts et méthodes des processus ponctuels, on développera les notions nécessaires à une analyse exploratoire de l'intensité et de la dépendance d'une configuration de points, en mettant l'accent sur l'étude des processus marqués. Pour finir, nous étudierons la modélisation de la régression géographiquement pondérée (RGP) appliquée aux données ponctuelles. Ce chapitre ne vise pas à offrir une revue exhaustive des théories et méthodes disponibles, mais plutôt à présenter les concepts fondamentaux et pertinents pour répondre à la question de recherche liée aux données de la CRC. Le tableau 2.1 liste les différents jeux de données qui seront utilisés à titre d'exemples illustratifs dans ce chapitre.

Jeux de données	Paquetages R	Descriptions
swedishpines	*spatstat*	Les données indiquent l'emplacement des jeunes arbres dans une forêt suédoise (Strand, 1972) et Ripley (1981).
anemones	*spatstat*	Les données présentent les localisation et les diamètres d'anémones de mer situées sur un rocher à Quiberon, France (Kooijman, 1979).
finpines	*spatstat*	Les données enregistrent les emplacements de 126 pins dans une forêt finlandaise, leurs hauteurs et leurs diamètres (Penttinen et collab., 1992).
EWHP	*GWmodel*	Les données présentent le prix de vente et les caractéristiques de maisons situées en Angleterre et au Pays de Galles (Fotheringham et collab., 2003). Les variables sont détaillées au tableau 2.3.

TABLEAU 2.1 – Liste des jeux de données utilisés dans le chapitre 2.

2.1 Introduction au concept des processus ponctuels

Un processus ponctuel est un mécanisme aléatoire produisant une configuration de points. Usuellement, les processus ponctuels sont notés avec des lettres majuscules (X, Y, etc.). Pour des applications statistiques, on peut généralement supposer que le nombre de points dans le processus est fini. Un processus ponctuel fini est un mécanisme aléatoire pour lequel chaque résultat possible est une configuration de points avec un nombre fini de points tel que, pour chaque région B, le nombre $n(X \cap B)$ de points à l'intérieur de B est une variable aléatoire bien définie. Ces deux conditions suffisent généralement à supporter les théories statistiques pour l'analyse des configurations des points. S'il existe de nombreux types de processus ponctuels (entre-autres Cox, Gibbs, Neymann-Scott, binomial), nous présenterons ici les processus de Poisson homogène et non-homogène qui sont les plus courants.

2.1.1 Processus de Poisson homogène *(Complete Spatial Randomness, CSR)*

Un processus de poisson homogène est un processus ponctuel permettant de générer des distributions spatiales complètement aléatoires. Il est caractérisé par deux propriétés clés :

- l'homogénéité : les points n'ont aucune « préférence » pour une localisation particulière ;
- l'indépendance : les réalisations dans une région de l'espace n'ont aucune influence sur les réalisations dans une autre région.

L'homogénéité implique que le nombre des points attendus dans la région B, noté $n(X \cap B)$, doit être proportionnel à sa surface, soit $E[n(X \cap B)] = \lambda |B|$, où λ est une constante qui correspond au nombre moyen de points par unité de surface.

De plus, ces deux propriétés d'homogénéité et d'indépendance impliquent que la distribution des points suit une loi de Poisson. Ainsi, le nombre $n(X \cap B)$ de points aléatoires situés dans la région B suit une distribution de Poisson. Dans la figure 2.1, on peut ainsi remarquer que le nombre de points pour chaque configuration est différent, alors qu'ils ont été simulés avec le même processus de Poisson d'intensité $\lambda = 10$.

Le processus de Poisson homogène est un modèle réaliste pour certains phénomènes physiques comme la radioactivité et les événements rares ou extrêmes. Il sert de référence auquel d'autres modèles peuvent être comparés. Ainsi, dans de nombreux tests statistiques, le processus de Poisson homogène sert d'hypothèse nulle.

2.1.2 Processus de Poisson non-homogène

Le modèle le plus important pour de nombreuses applications pratiques est le processus de Poisson non-homogène, qui est une modification du processus de Poisson homogène et pour lequel la densité moyenne des points varie dans l'espace.

Le processus de Poisson non-homogène est défini par les propriétés suivantes :

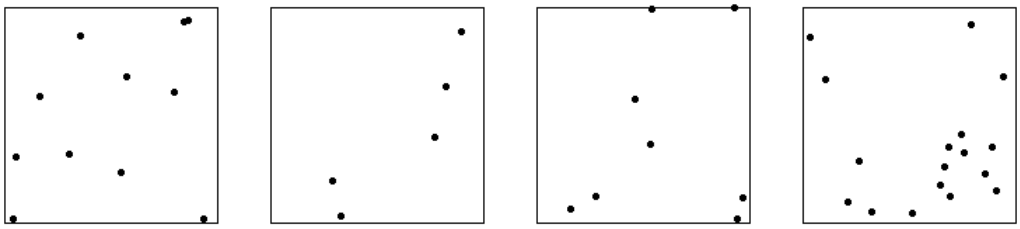


FIGURE 2.1 – Processus de Poisson homogène : simulation de 4 configurations de points avec une intensité de 10 points par unité de surface. Fonction `rpoispp()` de *spatstat*.

- Fonction d'intensité : le nombre de points attendus dans une régions B est variable et est défini par la fonction d'intensité : $E[n(X \cap B)] = \mu(B) = \int_B \lambda(x)dx$;
- Indépendance : si l'espace est divisé en régions qui ne se superposent pas, les configurations aléatoires dans ces régions sont indépendantes les unes des autres.

De plus, comme pour le processus de Poisson homogène, le nombre aléatoire de points dans une région donnée suit une distribution de Poisson (le paramètre de la loi devient $\mu(B)$ et non plus $\lambda|B|$). Dans ce cas où l'intensité n'est pas constante, elle peut être estimée avec des méthodes non paramétriques utilisées pour l'estimation de la densité.

2.2 Analyses exploratoires

Cette section a pour objectif de présenter différentes méthodes permettant d'effectuer des analyses exploratoires pour une configuration de points. Il s'agira dans un premier temps d'analyser l'intensité du processus, puis dans un deuxième temps d'étudier la dépendance entre les points.

2.2.1 Intensité

L'étude de l'intensité (ou moment d'ordre un) est l'une des premières et des plus importantes étapes dans l'analyse des données ponctuelles. L'intensité est une caractéris-

tique descriptive d'une configuration de points et représente le nombre de points attendus par unité de surface. L'intensité peut être constante (considérée comme « uniforme » ou « homogène ») ou bien elle peut varier d'un endroit à un autre (« hétérogène », « non-uniforme » ou « non-homogène »).

Par rapport à d'autres propriétés des processus ponctuels, l'intensité ne requiert pas beaucoup d'hypothèses de modélisation. On peut supposer une intensité homogène (constante) s'il y a une justification théorique qui l'explique. Par exemple, une grande partie de la cosmologie moderne suppose que l'univers est homogène à des échelles suffisamment grandes (Martinez et Saar, 2001). Cependant, cette hypothèse d'homogénéité peut être inappropriée pour de nombreuses configurations de points si l'intensité varie dans l'espace. Dans certains cas, la question principale de recherche est de savoir si l'intensité est homogène ou hétérogène. Par exemple, l'hétérogénéité de l'intensité d'une configuration de points peut démontrer la préférence ou l'évitement de certains animaux pour certains types d'habitat. Lorsque l'intensité varie spatialement, celle-ci est une fonction de la localisation spatiale, et il existe des méthodes statistiques pour estimer cette fonction à partir des données.

Intensité homogène, constante, uniforme

Un processus ponctuel X est défini comme ayant une intensité homogène si, pour chaque sous-région B d'un espace en deux dimensions, le nombre de points attendus de X dans B est proportionnel à la superficie de B :

$$E[n(X \cap B)] = \lambda |B|,$$

où λ est une constante appelée « intensité ». Cette intensité λ représente le nombre de points attendus par unité de surface, par exemple, deux points par mètre carré. La valeur de l'intensité dépend de l'unité de mesure de la surface. Ainsi, une intensité de deux points par mètre carré est alors équivalente à une intensité de 200 000 points par kilomètre carré.

En statistique de base, la moyenne de l'échantillon d'un jeu de données est une estimation sans biais de la moyenne de la population. De manière similaire pour les processus

ponctuels, la densité empirique des points $\bar{\lambda}$ est une estimation non-biaisée de la vraie intensité λ, en supposant que le processus ponctuel ait une intensité homogène. Ainsi, on a :

$$\bar{\lambda} = \frac{n(\mathbf{x})}{|W|},$$

où $\mathbf{x}$ est le jeu de données observé dans une fenêtre W et $n(x)$ est le nombre de points dans $\mathbf{x}$.

Intensité hétérogène

En général, l'intensité d'un processus ponctuel ne sera pas constante et variera d'un endroit à l'autre de la zone d'étude. Supposons que le nombre attendu de points situés dans une petite zone du autour d'un endroit u est égal à $\lambda(u)du$, où $\lambda(u)$ est la fonction d'intensité du processus. Alors, cette fonction d'intensité satisfait : $E[N(X \cap B)] = \int_B \lambda(u)du$ pour toutes les régions B.

Cependant, la fonction d'intensité pour certains processus ponctuels n'existe pas. Par exemple, les épicentres de tremblements de terre sont souvent concentrés très précisément le long d'une faille, à la frontière d'une plaque tectonique. On parle alors d'une « mesure d'intensité » Λ, définie par $\Lambda(B) = E[N(X \cap B)]$ pour chaque $B \subset \mathbf{R}^2$. S'il est possible que l'intensité d'un processus ponctuel ne soit pas homogène, alors on peut l'estimer avec des techniques non paramétriques comme la méthode des quadrats ou l'estimation de la densité par noyau.

Estimation de l'intensité par la méthode des quadrats

Une manière simple de vérifier l'hétérogénéité d'un processus est de regarder si les régions de même superficie contiennent environ le même nombre de points (comme cela devrait être le cas pour un processus d'intensité homogène). Dans la méthode des quadrats, la fenêtre d'observation W est divisée en sous-régions $B_1, ..., B_m$ appelées « quadrats ». Pour des raisons de simplicité, on suppose que ces sous-régions sont de même superficie. Pour chaque quadrat (sous-région), on compte le nombre de points situés à l'intérieur :

$n_j = n(\mathbf{x} \cap B_j)$ pour $j = 1, \ldots, m$. Ainsi, si l'intensité est homogène, les nombres de points dans chaque quadrat devraient être relativement égaux.

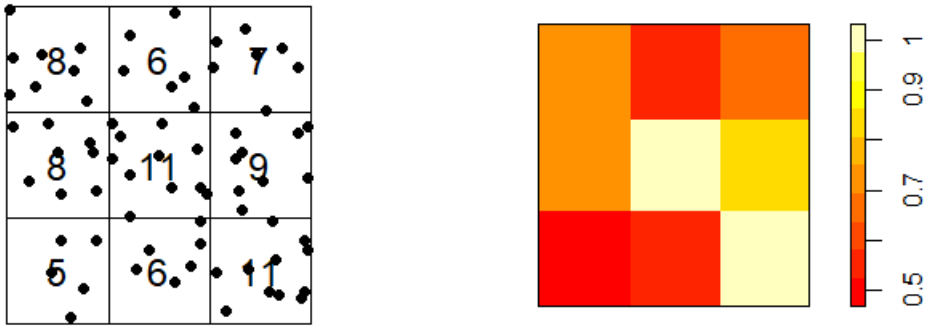


FIGURE 2.2 – Gauche : Dénombrements par la méthode des quadrats pour le jeu de données *swedishpines* . Droite : Mesure d'intensité (points par mètre carré).

Il est important de noter que lorsqu'on choisit la taille des quadrats, il y a un compromis à faire entre biais et variabilité. En effet, de plus grands quadrats réduisent l'erreur relative des dénombrements n_j mais aussi effacent la variation spatiale de l'intensité dans chaque quadrat. Cependant, les quadrats peuvent en général avoir des formes et des superficies inégales. Au lieu de définir des quadrats en carrés, il est possible de les définir avec une forme hexagonale comme sur la figure 2.3. Lorsque les quadrats n'ont pas la même superficie, les dénombrements de points dans ces quadrats ne peuvent pas être comparés directement, mais sous l'hypothèse d'homogénéité, le nombre attendu dans chaque quadrat reste proportionnel à la superficie du quadrat. Ainsi, la densité empirique $\bar{\lambda}$ de chaque quadrat reste un estimateur non biaisé de l'intensité homogène λ.

Test d'homogénéité selon la méthode des quadrats

Il est possible d'effectuer un test statistique afin d'évaluer si l'intensité d'un processus

ponctuel est homogène ou non. Pour des raisons de praticité, on supposera provisoirement que le processus ponctuel est un processus de Poisson, ainsi :

H_0 : L'intensité est homogène (Processus de Poisson homogène),

H_1 : L'intensité est hétérogène (Processus de Poisson non-homogène).

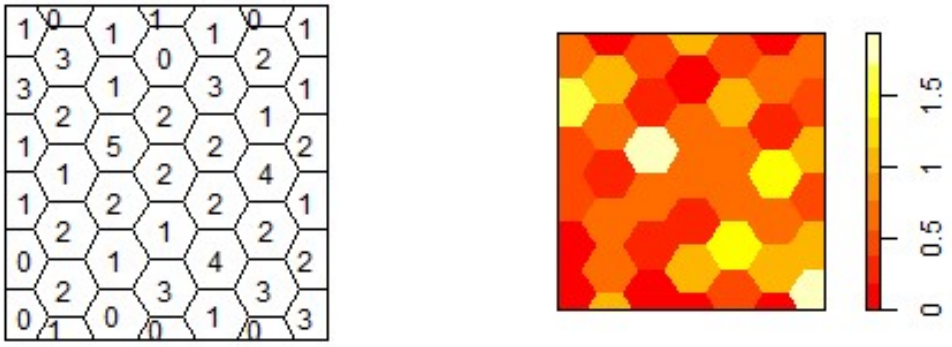


FIGURE 2.3 – Gauche : Dénombrements par la méthode des quadrats hexagonaux pour le jeu de données *swedishpines*. Droite : Mesure d'intensité (points par mètre carré).

Comme précédemment, la fenêtre W est divisée en m quadrats $B_1, \ldots, B_m$ et on compte le nombre $n_1, \ldots, n_m$ de points présents dans chacun des quadrats. Si l'hypothèse nulle est vraie, alors les n_j proviennent d'une variable aléatoire de Poisson avec des valeurs espérées $\mu_j = \lambda a_j$, où λ est la vraie intensité (qui est inconnue) et a_j est la superficie de B_j. Sous l'hypothèse nulle, la distribution de la statistique de test $X^2 = (m/n)\sum_j (n_j - n/m)^2$ suit une loi de χ^2 avec $m-1$ degrés de liberté.

Pour le jeu de données *swedishpines*, avec une valeur-p de 0.42, ce test d'hypothèse ne nous permet pas de rejeter l'hypothèse nulle d'une distribution spatiale complètement aléatoire. Cette configuration de points est en effet considérée comme ayant une intensité homogène (Strauss, 1975; Ripley, 1981).

En incluant une ou plusieurs covariables qui ont potentiellement une influence sur l'intensité, il est possible d'effectuer des tests d'hypothèses plus puissants. L'hypothèse

nulle ne reposerait alors plus sur une distribution spatiale complètement aléatoire, mais serait basée sur la distribution de la covariable d'intérêt.

Si la méthode des quadrats est simple à appliquer, elle représente en revanche certains défauts (Baddeley et collab., 2015). Les résultats sont en effet sensibles à la spécification des quadrats (forme et taille) car elle sous-entend que la densité est la même dans tous les sous-ensembles de la région étudiée. De plus, le test statistique présenté plus haut ne permet pas d'obtenir beaucoup d'information sur le processus étudié car l'hypothèse alternative représente uniquement le fait que le processus ponctuel n'est pas un processus de Poisson homogène. Or un processus de Poisson homogène est caractérisé par son intensité uniforme mais aussi par l'indépendance entre ses points. En effectuant ce test, on suppose ainsi que les points ne sont pas dépendants entre-deux, cependant cette supposition peut être considérée comme naïve car l'indépendance des observations est souvent discutable.

Estimation de la densité par la méthode du noyau ou lissage spatial de la fonction d'intensité

La fonction d'intensité théorique en un point x est obtenue en calculant la moyenne des points observés par unité de surface sur des voisinages contenant x de plus en plus petits.

Le lissage spatial n'estime pas directement la fonction d'intensité mais une version lissée obtenue par convolution avec un noyau K. La fonction d'intensité $\lambda(u)$ d'une configuration de points peut ainsi être estimée grâce à la méthode non-paramétrique d'estimation par noyau (*kernel estimation*).

Un estimateur à noyau est composé d'une somme de n fonctions (une pour chaque observation) souvent accompagnées d'une correction liée aux effets de bords. Pour chaque emplacement spatial u dans une fenêtre W, les estimateurs de noyau de la fonction d'intensité les plus courants sont :

- sans correction :

$$\tilde{\lambda}^{(0)}(u) = \sum_{i=1}^{n} K(u - x_i),$$

- correction uniforme :

$$\tilde{\lambda}(u) = \frac{1}{e(u)} \sum_{i=1}^{n} K(u - x_i),$$

- correction de Diggle :

$$\tilde{\lambda}(u) = \frac{1}{e(x_i)} \sum_{i=1}^{n} K(u - x_i),$$

où $K(u)$ est une fonction appelée noyau et où $e(u) = \int_W K(u-v)dv$ est une correction due aux effets de bords. En dehors de la fenêtre d'observation W, l'estimation de l'intensité est égale à 0. L'écart-type du noyau est appelé « bande passante » (*bandwidth*) et permet de contrôler le degré de lissage de l'estimation de la fonction d'intensité. Dans d'autres termes, le noyau décrit la manière dont le voisinage est appréhendé et la bande passante est un paramètre qui permet de quantifier la « taille » du voisinage. Lorsque le paramètre de bande passante est pertinent, les estimations obtenues sont statistiquement robustes et permettent d'établir si la fonction d'intensité est homogène ou hétérogène dans l'espace.

Traitement des effets de bord

En raison des effets de bords, l'estimation non corrigée présente une fort biais négatif aux endroits proches des frontières de la fenêtre d'observation W. Cette estimation doit donc être uniquement utilisée dans les rares situations où il n'y a pas d'effets de bord. La correction uniforme et la correction de Diggle (Diggle, 1985) sont ainsi conçues pour compenser l'effet de bord se produisant lorsqu'un processus ponctuel est observé à l'intérieur d'une fenêtre W. Elles se distinguent par leur manière d'appréhender l'extérieur de la fenêtre d'observation W, ainsi que par leur rapidité d'exécution (Baddeley et collab., 2015).

Ainsi, lorsque la fenêtre d'observation W est indépendante du processus sous-jacent, la correction uniforme assure la continuité de l'intensité entre l'intérieur de la fenêtre et l'extérieur. L'estimateur utilisant la correction uniforme est alors sans biais lorsque

l'intensité réelle est homogène. À l'inverse, si l'intensité en dehors de W est jugée nulle, la correction de Diggle est davantage appropriée mais nécessite un temps de calcul plus élevé.

Choix du noyau

Diverses fonctions noyau produisent des différences subtiles dans la forme de la surface lissée. Ces différences sont majoritairement dues au fait que, selon le noyau, on accorde une place plus ou moins importante au critère de distance. Généralement, le choix du noyau impacte peu les résultats du lissage.

Voici les noyaux les plus courants et leurs spécificités :

- Noyau Gaussien : il prend en compte l'ensemble des points de la zone d'étude (contrairement aux autres noyaux). $K_h(x) = \frac{1}{2\pi} e^{||\frac{x}{h}||^2}$.
- Noyau quadratique : il donne un poids plus élevé aux points les plus proches qu'aux points les plus éloignés. La décroissance est graduelle mais le noyau s'annule au-delà du rayon de lissage. $K_h(x) = \frac{9}{16} 1_{||x||<h}(1 - ||\frac{x}{h}||^2)^2$.
- Noyau uniforme : il pondère chaque point dans un cercle de rayon r (rayon de lissage) de manière égale. Il n'y a donc pas de pondération selon la distance.

Choix de la bande passante (*bandwidth*)

La bande passante (aussi appelé rayon de lissage) du noyau contrôle le degré de lissage de l'estimation de la fonction d'intensité. Nous pouvons voir sur la figure 2.4 qu'une petite valeur pour la bande passante produit une surface d'intensité plutôt irrégulière, alors qu'une grande valeur semble lisser trop fortement l'intensité. Le choix de la bande passante résulte d'un arbitrage biais-variance entre la précision spatiale de l'analyse et sa qualité statistique. Ainsi, plus la bande passante est grande, plus le biais augmente et ainsi plus le nombre de points participant au calcul des estimations locales augmente ; donc la variance diminue. Par conséquent, comme dans toutes les méthodes non paramétriques, si le choix du noyau a un impact limité, celui de la bande passante est, lui, extrêmement important. Différentes méthodes sont disponibles pour proposer automatiquement une

bande passante qui minimise un critère d'erreur. Certaines de ces méthodes utilisent la validation croisée, et supposent que la distribution de points observée suit une distribution de Poisson pour pouvoir estimer une bande passante optimale.

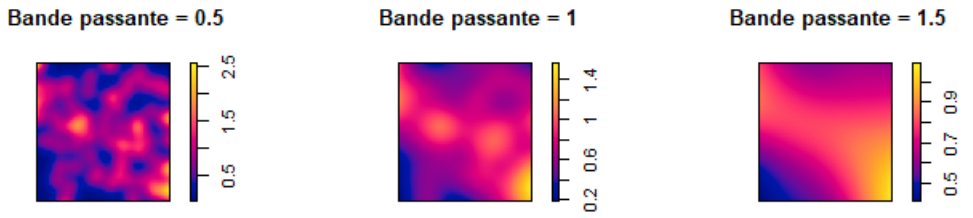


FIGURE 2.4 – Estimation de la densité du jeu de donnée *swedishines* (disponible dans *spatstat*) selon différentes valeurs de bande passante pour un noyau Gaussien avec une correction « uniforme ».

2.2.2 Dépendance et corrélation

Définition et concepts de base

Lorsque l'on étudie une configuration de points, un des objectifs récurrents est de pouvoir déterminer si les points ont été placés indépendamment les uns des autres, ou s'ils présentent une quelconque dépendance entre eux. Dans le cas où ces points n'auraient pas été distribués au hasard, on s'intéresse alors à la nature de cette interdépendance : est-elle répulsive ou agrégative ? La figure 2.5 présente les trois configurations de points typiques :

- **Configuration agrégée : les points ont tendance à se rassembler**
  Il existe une interaction entre les points, ils s'attirent et créent des agrégats (*clusters*). On peut alors détecter une concentration géographique. Par exemple, certains types de magasins peuvent être concentrés uniquement dans les centres-villes ou même dans des secteurs spécifiques. Les magasins de luxe seront par exemple souvent regroupés dans les mêmes quartiers voire dans les mêmes rues.

- **Configuration régulière (répulsive) : les points ont tendance à s'éviter entre eux**

  Avec cette configuration, les points sont plus régulièrement espacés qu'ils ne le seraient sous une distribution complètement aléatoire. On peut par exemple penser à la répartition des arbres le long d'une rue ou dans un verger : ils se repoussent et créent une distribution de points dispersés.

- **Configuration complètement aléatoire : les points n'exercent aucune interaction entre eux**

  Si toutes les configurations de points, en tant que réalisation d'un processus ponctuel, sont aléatoires, celle-ci est qualifiée de « complètement » aléatoire car les points sont localisés partout avec la même probabilité et indépendamment les uns des autres. Cette configuration correspond au Processus de Poisson homogène *(CSR)* qui, comme présenté dans les sections précédentes, est au cœur de la théorie des processus ponctuels. Dans la pratique, il est souvent difficile de détecter une telle configuration à l'œil nu, et seule l'utilisation d'indicateurs permet de juger si la distribution observée s'écarte de manière significative d'une distribution complètement aléatoire.

La corrélation (ou plus généralement la covariance) est un outil statistique incontournable pour mesurer la dépendance. Cependant, une certaine prudence est indispensable lors de l'étude et de l'analyse des résultats de corrélation (Floch et collab., 2018). En effet, une mesure précise de la corrélation exige une estimation fidèle de la moyenne, ce qui, dans le cas des configurations de points, correspond à une bonne estimation de l'intensité du processus. Sur la figure 2.5, les mêmes structures agrégées ou régulières peuvent être obtenues grâce à un processus de Poisson non-homogène (dans lequel l'intensité du processus varie dans l'espace) mais où les points sont indépendants les uns des autres. En étudiant la localisation d'entreprises, Ellison et Glaeser (1997) ont démontré que les effets de certains avantages naturels (qui induisent une intensité plus forte) ne sont pas différenciables de ceux d'externalités positives (qui génèrent de l'agrégation). Ainsi, la confusion

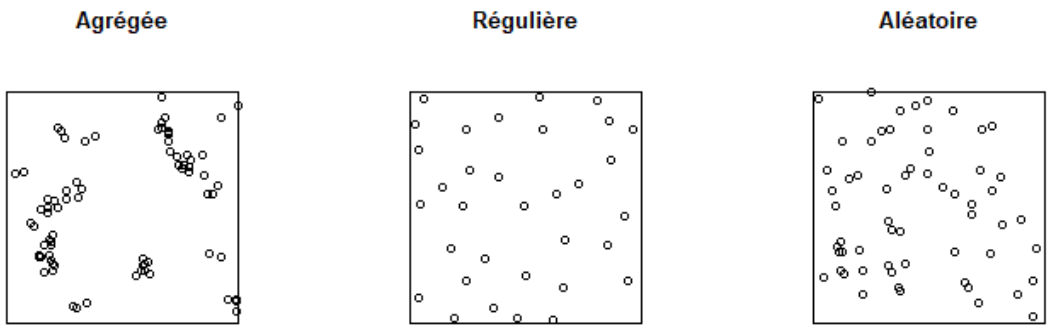


FIGURE 2.5 – Représentation des trois configurations-types : agrégée, régulière et aléatoire. Simulées avec les fonctions `rMatClust()` *[agrégée]*, `rMaternII()` *[régulière]* et `rpoispp()` *[aléatoire]* de *spatstat*.

entre l'intensité et la corrélation peut aussi concerner les processus en eux-mêmes.

Une autre mise en garde concerne l'homogénéité car les méthodes développées consistent souvent à tester l'existence d'une agrégation ou d'une répulsion en assumant au préalable un processus homogène. Il s'agit alors de tester une configuration de points avec comme hypothèse nulle une distribution complètement aléatoire *(CSR)*. La fonction *K* de Ripley est une méthode incontournable pour cela. Cependant, une telle hypothèse nulle *(CSR)* peut parfois ne pas être adéquate et on privilégiera d'autres variantes de cette méthode. Cette section aura pour but de présenter les différentes méthodes principalement utilisées pour identifier des corrélations et caractériser correctement une distribution de points.

La fonction *K* de Ripley (et ses variantes)

Aussi connue sous le nom de « fonction de Ripley », cette méthode proposée par Ripley (1976) est la plus utilisée pour étudier la corrélation spatiale dans les processus ponctuels. Elle permet d'estimer le nombre moyen de voisins rapporté à l'intensité du processus. C'est une fonction cumulative, donnant le nombre moyen de voisins présents à une distance inférieure à un rayon *r* pour chaque point, et standardisée par l'intensité du

processus $\frac{n}{|W|}$ qui est supposé homogène. Son estimateur $\widehat{K}(r)$ se calcule de cette manière :

$$\widehat{K}(r) = \frac{|W|}{n(n-1)} \sum_i \sum_{j \neq i} 1\{\|x_i - x_j\| \leq r\} c(x_i, x_j; r),$$

où n est le nombre total de points sur la fenêtre d'observation et donc $n(n-1)$ est le nombre total de paires de points, $1\{\|x_i - x_j\| \leq r\}$ est une indicatrice qui vaut 1 si les points i et j sont à une distance plus petite ou égale à r et 0 sinon. La somme de ces indicatrices représente donc le nombre de fois où les points i et j sont à une distance inférieure ou égale à r. Le terme $c(x_i, x_j; r)$ correspond à la correction des effets de bord et W à l'aire d'étude. De manière concrète, on balaye toutes les distances r et pour chacune d'entre-elles, on calcule la valeur de la fonction K pour étudier le voisinage des points. On procède de cette manière :

1. Pour chaque point et pour chaque distance r, on compte le nombre de voisins de ce point qui sont situés sur le disque de rayon r,

2. Ensuite, on calcule le nombre moyen de voisins pour chaque distance r (en tenant comptes d'éventuels effets de bords),

3. On compare ensuite ces résultats à ceux obtenus sous l'hypothèse d'une distribution homogène *(CSR)* qui constitue une valeur de référence.

L'objectif principal de cette méthode est alors de détecter s'il existe un écart significatif entre les estimations du nombre de voisins observés et attendus (valeur de référence). La standardisation et la correction des effets de bord permettent de comparer les résultats de différentes configurations de points, qui ont été observées dans des fenêtres différentes et avec avec un nombre de points différents.

La figure 2.6 présente les courbes des fonctions K associées aux processus de la figure 2.5 vue précédemment. Chaque rayon r est représenté en abscisse et la valeur de la fonction K estimée à cette distance est représentée en ordonnée. La fonction K estimée est représentée par la courbe noire, et la courbe rouge (pointillés) représente la valeur de référence πr^2pour tous les rayons r.

L'interprétation des résultats se fait de la manière suivante :

- **En présence d'un processus agrégé :** la courbe de la fonction $\hat{K}$ est située au-dessus de celle de référence, car il y a en moyenne plus de points dans un rayon r autour des points de la configuration étudiée par rapport aux points attendus pour une distribution aléatoire ($\hat{K}(r) > \hat{K}_{pois}(r)$).

- **En présence d'un processus régulier :** la courbe de la fonction $\hat{K}$ est située au-dessous de celle de référence, car s'ils se repoussent, ils ont moins de voisins situés dans un rayon r que pour un processus complètement aléatoire ($\hat{K}(r) < \hat{K}_{pois}(r)$).

- **En présence d'un processus complètement aléatoire :** la courbe de la fonction $\hat{K}$ est très similaire à celle de référence. Sur le graphique de la figure 2.6, les deux courbes sont extrêmes rapprochées jusqu'à un rayon de 0,1, où on distingue une déviation très légère. Selon les configurations, il se peut que les courbes ne soient pas similaires, mais elles resteront tout de même très semblables.

Il est important de souligner que la fonction K est définie sous l'hypothèse de station-

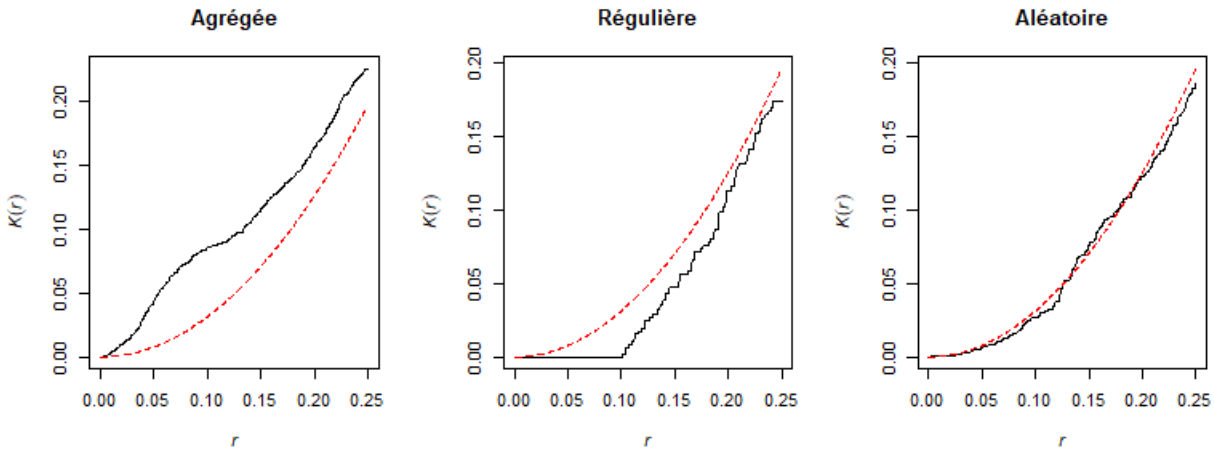


FIGURE 2.6 – Les fonctions K des configurations-types (présentées à la figure 2.5). Calculées à l'aide de la fonction `Kest()` du paquetage *spatstat*. La fonction K estimée est représentée par la courbe noire, et la courbe rouge (pointillés) représente la valeur de référence πr^2pour tous les rayons r.

narité. Ainsi, pour un processus de poisson non homogène, l'écart entre l'estimation de la fonction K et celle de référence peut être dû à la variation d'intensité, et non pas à un phénomène d'attraction. La figure 2.7 illustre ce phénomène. La configuration de points a été simulée avec une intensité plus forte à droite de la fenêtre, mais sans attraction entre les points. On observe alors que l'analyse basée sur la fonction K (graphique du milieu) est biaisée car les résultats montrent un processus agrégé, or ce n'est pas le cas. De plus, comme en statistique classique, la corrélation n'entraîne pas la causalité et une absence de corrélation ne signifie pas non plus forcément une indépendance.

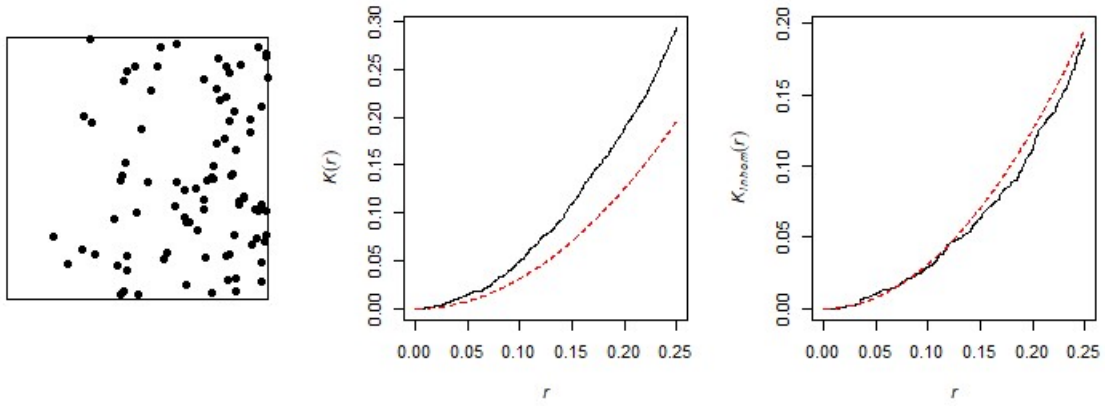


FIGURE 2.7 – Configuration de points avec une intensité non homogène (gauche), la courbe de la fonction K associée (milieu) et la courbe de la fonction K_{inhom} (à droite). Les fonctions $\hat{K}$ estimées sont représentées par la courbe noire, et la courbe rouge (pointillés) représente la valeur de référence πr^2pour tous les rayons r. La configuration de points a été simulée à l'aide de la fonction `rpoispp()` de *spatstat*.

On retrouve fréquemment deux variantes de la fonction K : La fonction L de Besag (1977) et la fonction D de Diggle et Chetwynd (1991). Si ces méthodes (K, L et D) sont particulièrement intéressantes, c'est parce qu'elles analysent l'espace étudié en parcourant toutes les distances, et ne retiennent pas uniquement qu'un seul ou quelques niveaux géographiques. Ainsi, ces méthodes sont les seules à permettre de détecter exactement les distances où il y a des phénomènes d'attractions ou de dispersions, sans biais d'échelle lié à un zonage prédéfini. Des structures assez complexes peuvent alors être détectées avec,

au sein d'une même configuration de points, de la répulsion pour certaines distances mais aussi de l'agrégation pour d'autres.

La fonction L est une transformation de la fonction K proposée par Besag (1977) et qui est définie par $L = (K(r)/\pi)^{\frac{1}{2}}$ et équivaut à $L_{pois}(r)$ lorsque le processus est aléatoire.

La fonction D (Diggle et Chetwynd, 1991) permet de confronter les distributions de deux sous-populations et ainsi de prendre en compte la non-homogénéité de l'espace. En effet, les analyses des fonctions K et L ne sont pertinentes que lorsque l'hypothèse d'homogénéité est vérifiée, ce qui n'est pas souvent le cas. Initialement, la fonction D a été élaborée dans un contexte épidémiologique pour comparer la concentration des enfants atteints d'une maladie rare et celle des enfants sains dans une même région (les sous-populations ont donc la même répartition spatiale). Elle est ainsi définie par $D(r) = K_{\text{cas}}(r) - K_{\text{controles}}(r)$ et permettra de mettre en évidence si les cas sont plus agrégés que les contrôles, ou inversement.

Une autre variante de la fonction K permettant de prendre en compte la non-homogénéité de l'espace est la fonction K_{inhom} (Baddeley et collab., 2000), définie de cette manière :

$$\widehat{K}_{\text{inhom}}(r) = \frac{1}{D}\sum_i \sum_{j \neq i} 1\frac{\{\|x_i - x_j\| \leq r\}}{\widehat{\lambda}(x_i)\widehat{\lambda}(x_j)} e(x_i, x_j; r),$$

avec $D = \frac{1}{|W|}\sum_i \frac{1}{\lambda(x_i)}$ et où $\widehat{\lambda}(x_i)$ est l'estimation de l'intensité du processus autour du point i. Elle utilise ainsi les valeurs estimées de l'intensité mais nécessite une bonne estimation des densités locales (par la méthode des noyaux), ce qui est parfois complexe dans la réalité.

La figure 2.7 présente (à droite) la fonction K_{inhom} pour la configuration étudiée précédemment, qui est très similaire à la courbe de référence. On retrouve donc un résultat cohérent avec le processus simulé (pas d'attraction entre les points mais une intensité non homogène). Pour cet exemple, la fonction K est donc inadaptée à l'étude de la dépendance entre les points, et on lui préférera la fonction K_{inhom}.

2.3 Processus marqués

Comme introduit au chapitre 1, il est possible que chaque point d'une configuration soit accompagné d'un ou plusieurs attributs qui les caractérisent et que l'on appelle des « marques ». Ces marques peuvent être catégorielles (qualitatives) comme par exemple les différents types d'espèces, ou bien numériques (quantitatives) comme l'âge ou encore la taille des arbres si la zone d'étude est une forêt. Par soucis de cohérence avec l'application des méthodes appliquées aux données de la CRC, on s'intéressera ici uniquement à l'analyse des marques quantitatives.

2.3.1 Analyse exploratoire

En présence d'une configuration de points avec des marques numériques, une des premières étapes de l'analyse exploratoire est d'effectuer une analyse descriptive pour chacune des marques. On peut donc étudier chaque marque à l'aide des outils standards d'analyse de données (histogrammes, nuages de points, boîtes à moustaches, etc.). L'histogramme de la figure 2.8 représente le diamètre des anémones du jeu de données *anemones*. Les anémones avec un diamètre de 4 unités sont les plus représentées et la taille moyenne du diamètre est de 4.29 unités. Commune aux autres types de données, cette étape d'inspection permet de mieux connaître les données et de déceler de potentielles anomalies ou bien des tendances.

Sur le nuage de points de la figure 2.9, on a représenté le diamètre en fonction de la hauteur des arbres du jeu de données *finpines* (illustré à gauche de la figure). On remarque ainsi directement que la hauteur et le diamètre sont relativement proportionnels et que les diamètres ont des valeurs discrètes (contrairement aux hauteurs).

On peut aussi examiner une potentielle tendance spatiale des marques à l'aide de visualisations comme la figure 2.8 (gauche) qui représente à la fois la localisation des points mais aussi la marque associée (les différents diamètres). Ici, la représentation avec différentes tailles de points est relativement lisible car les valeurs de la marque sont discrètes et ne sont pas nombreuses. Dans le cas contraire, il peut être judicieux de les convertir en

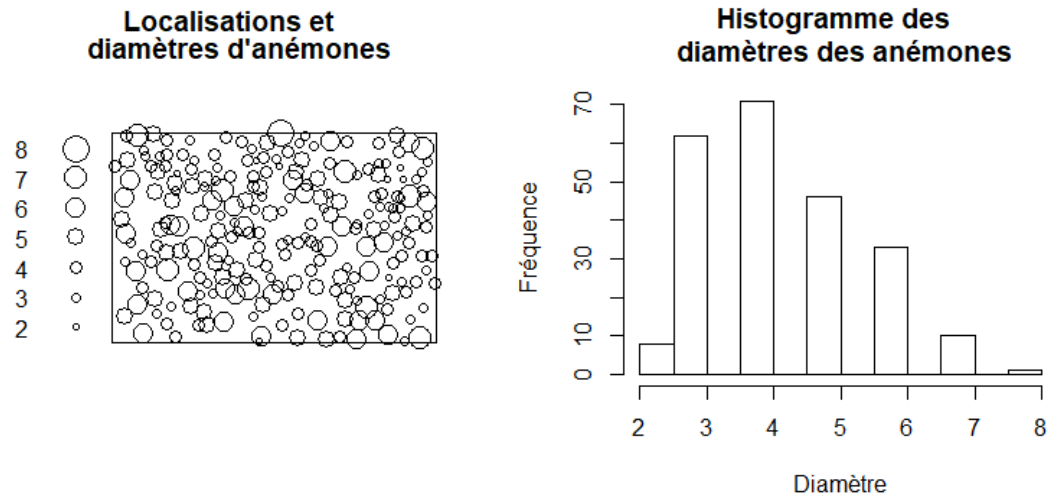


FIGURE 2.8 – Localisations et diamètres d'anémones de mer (jeu de données *anemones* de *spatstat*). Données recueillies par Kooijman (1979) sur un rocher à Quiberon en France.

catégories discrètes pour ensuite les visualiser avec différentes couleurs ou formes. Sur cette figure, il est cependant difficile d'établir une tendance spatiale de la marque car les différents diamètres semblent présents dans toutes les zones. D'autres outils seront alors plus fiables pour distinguer s'il y a une homogénéité (ou non) de la valeur moyenne de la marque dans la zone étudiée.

Étude de l'intensité de la marque

Le lisseur Nadaraya–Watson (Nadaraya, 1964) est une fonction spatiale permettant de lisser les marques des points et peut être considéré comme une estimation de la valeur moyenne de la marque variant dans l'espace. En incluant la correction de Diggle, il est défini de cette manière :

$$\tilde{m}^{(D)}(u) = \frac{\sum_i m_i \kappa(u - x_i)/e(x_i)}{\sum_i \kappa(u - x_i)/e(x_i)},$$

où m_i est la marque correspondante au point x_i, $e(x_i)$ est la correction des effets de bords et κ est un noyau de lissage. Lorsque la bande passante (ou rayon de lissage) est grande, la valeur de la fonction deviendra approximativement constante et égale à la valeur moyenne

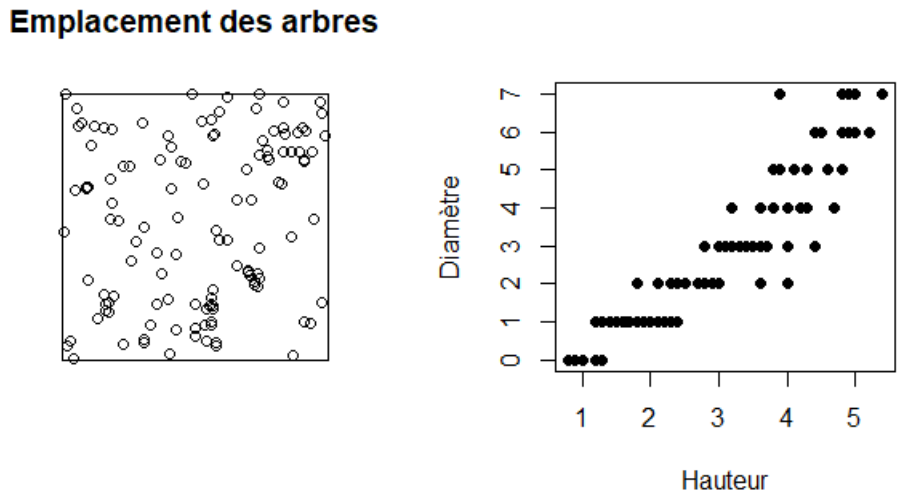


FIGURE 2.9 – Gauche : Emplacement de 126 arbres (jeunes pousses de pins) dans une forêt finlandaise (jeu de données *finpines* de *spatstat*). Droite : Nuage de points du diamètre en fonction de la hauteur des pins.

de la marque dans l'ensemble du jeu de données. À l'inverse, lorsque la bande passante tend vers 0, $\tilde{m}^{(D)}(u)$ tendra vers la valeur de la marque du point x_i le plus proche de l'emplacement u.

Il est important de noter que le lissage spatial d'une marque est différent du lissage spatial de la fonction d'intensité (présenté à la section 2.2.1). En effet, si le résultat du lissage de la fonction d'intensité à l'emplacement u donnera le nombre de points situés dans le voisinage de u, le lissage de la marque à l'emplacement u sera lui une moyenne des marques des points situés dans le voisinage de u.

La figure 2.10 présente le lissage des deux marques de la configuration de points *finpines* (présenté à la figure 2.9). Comme constaté grâce au nuage de points de l'analyse exploratoire du jeu de données (figure 2.9), on retrouve bien ici aussi la relation entre la hauteur et le diamètre des arbres. En effet, les valeurs (hauteur et diamètre) les plus élevées se retrouvent relativement aux mêmes endroits et on constate par exemple en bas à droite une présence d'arbres plus grands et plus larges qu'ailleurs dans la zone étudiée.

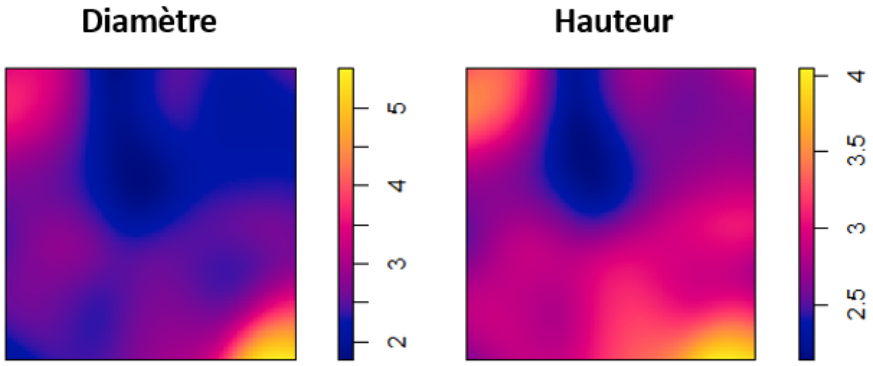


FIGURE 2.10 – Lissage des marques du jeu de données *finpines* de *spatstat*. Droite : Diamètres moyen des arbres (jeunes pousses de pin) variant dans l'espace (en centimètres). Gauche : Hauteur moyenne des arbres (jeunes pousses de pin) variant dans l'espace (en mètres).

Dépendance et corrélation

Il existe certains outils statistiques permettant d'effectuer des analyses exploratoires sur la dépendance des marques, notamment la fonction de corrélation de marque ou encore la fonction K pondérée par les marques.

La fonction de corrélation de marque (*mark correlation function*) d'un processus marqué stationnaire est une mesure de la dépendance entre les marques de deux points situés à une distance r l'un de l'autre (Stoyan et Stoyan, 1995) et est définie (de manière informelle) par :

$$\rho_f(r) = \frac{\mathrm{E}[f(M_1, M_2)]}{\mathrm{E}[f(M, M')]},$$

où M_1 et M_2 sont les marques des deux points séparés par la distance r et M, M' sont des marques aléatoires indépendantes et identiquement distribuées qui ont la même distribution que la marque d'un point choisi au hasard. Ici, f est une fonction qui retourne une valeur réelle non-négative. Les choix courants de f selon la nature des marques sont :

- marques continues : $f(m_1, m_2) = m_1 m_2$,

- marques catégorielles (processus multitypes) : $f(m_1, m_2) = 1\{m_1 = m_2\}$,
- marques représentant des angles ou des directions : $f(m_1, m_2) = \sin(m_1, m_2)$.

Ainsi, $\rho_f(r)$ peut prendre n'importe quelle valeur réelle non-négative et la valeur 1 suggère une absence de corrélation. L'interprétation des valeurs inférieures ou supérieures à 1 dépend du choix de la fonction f.

La figure 2.11 montre l'estimation de la fonction de corrélation pour la marque (diamètre) du jeu de données *anemones*. La ligne verte en pointillés représente la valeur 1 suggérant une absence de corrélation et les courbes rouges et noires représentent les fonctions de corrélation estimées $\hat{k}_{mm}$ avec respectivement les corrections « translate » et « isotropic » de la fonction `markcorr()`. L'écart entre les deux fonctions, résultant de la différence de correction, est très minime. Le graphique suggère une légère association négative entre les tailles des anémones voisines. En effet, la courbe estimée est située en dessous de la valeur 1 (pointillés verts) ce qui implique une corrélation négative. En revanche, cette association reste relativement faible dans la mesure où la courbe se rapproche de la valeur de référence aux alentours de $r = 15$.

La fonction K pondérée par les marques $K_f(r)$ d'un processus ponctuel marqué (Penttinen et collab., 1992) est une généralisation de la fonction K de Ripley (présentée à la section 2.2.2) et définie de cette manière :

$$K_f(r) = \frac{1}{E[f(M, M')]} E\left[\sum_{x_j \in X} f(m(u), m(x_j)) 1\{0 < ||u - x_j|| \leq r\} \middle| u \in X \right],$$

où $m(u)$ et $m(x_j)$ désignent les valeurs de la marque aux points u et x_j et M, M' sont des marques aléatoires indépendantes et identiquement distribuées qui ont la même distribution que les marques du processus ponctuel. Ainsi, sous l'hypothèse d'étiquetage aléatoire (*random labelling*), $K_f(r)$ est égale à $K(r)$, la fonction K classique. Pour effectuer des tests d'hypothèses, cette fonction cumulative est plus adaptée que la fonction de corrélation de marque. La figure 2.12 montre l'estimation de la fonction K pondérée par le diamètre des anémones. La courbe $\hat{K}$ estimée est similaire à la courbe de référence

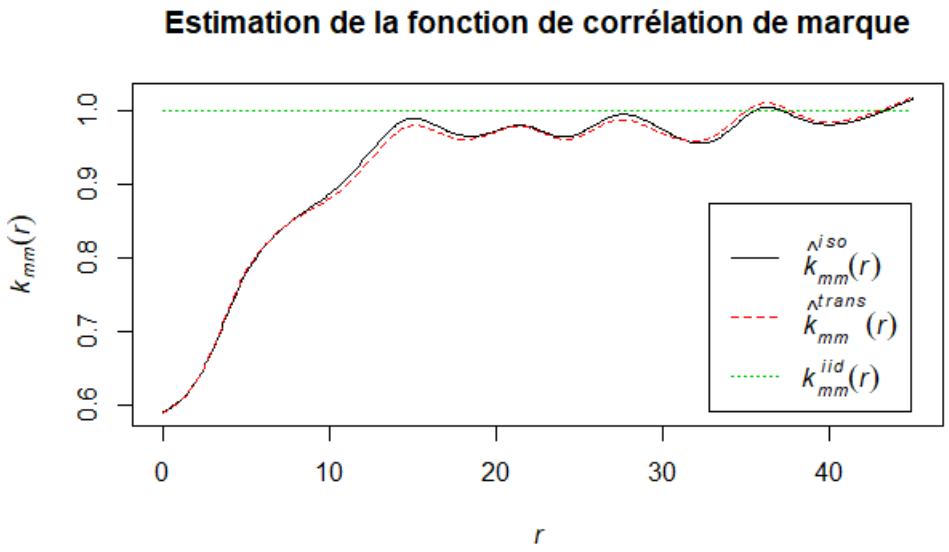


FIGURE 2.11 – Estimation de la fonction de corrélation de marque (diamètre) pour le jeu de données *anemones* disponible dans *spatstat*.

mais reste cependant toujours inférieure à celle-ci, ce qui est cohérent avec une légère association négative trouvée avec la fonction de corrélation de marque.

Une alternative aux fonctions de second ordre comme la fonction de corrélation des marques ou bien la fonction K pondérée est de **considérer uniquement les voisins les plus proches** avec par exemple le calcul de la corrélation classique entre la marque d'un point typique et la marque de son voisin le plus proche. Cette corrélation classique prend donc des valeurs entre -1 et 1, et les valeurs proches de ces deux extrémités indiquent une forte dépendance entre les marques. Pour le jeu de données *anémones*, cette corrélation est de 0,1 ce qui ne correspond pas à une forte corrélation entre les diamètres des anémones les plus proches.

2.4 Régression géographiquement pondérée (RGP)

Dans l'objectif de pouvoir présenter des théories et des méthodes pertinentes pour l'analyse des données de la Croix-Rouge, nous présentons ici la régression géographique-

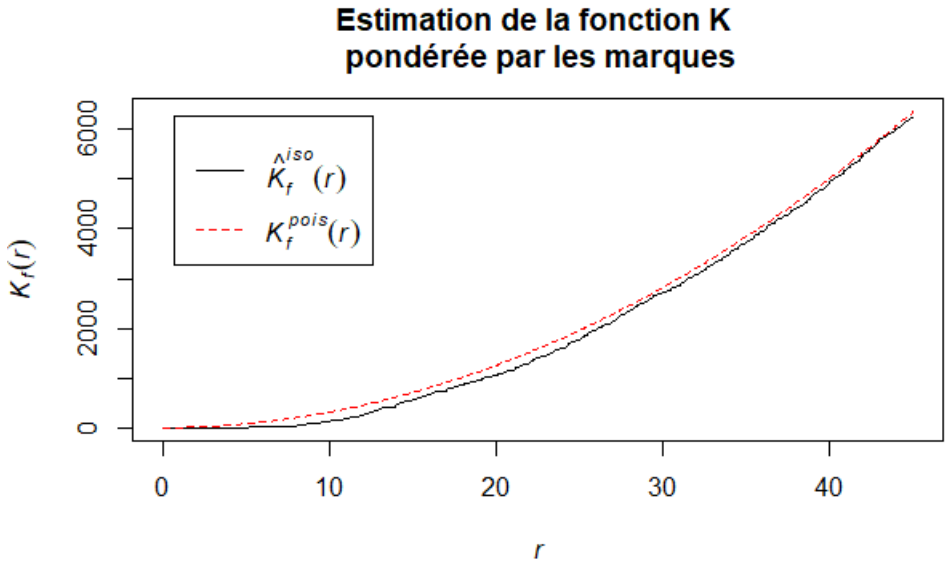


FIGURE 2.12 – Estimation de la fonction K pondérée par les marques (diamètre) pour le jeu de données *anemones* disponible dans *spatstat*.

ment pondérée comme méthode de modélisation appliquée aux données ponctuelles avec marques.

La régression géographiquement pondérée (*Geographically Weighted Regression, GWR*) a été introduite par Brunsdon et collab. (1996) pour décrire une famille de modèles de régression dont les coefficients β peuvent varier dans l'espace. Contrairement aux modèles classiques estimés sur l'ensemble d'un territoire, la RGP permet de comprendre les variations locales en estimant plusieurs modèles locaux qui constitueront ensuite un modèle global final.

La régression linéaire classique modélise la variable dépendante y comme une fonction linéaire des variables explicatives $x_1, ..., x_p$ et est définie de cette manière :

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \varepsilon_i,$$

avec $\beta_0, ..., \beta_p$ les paramètres de la régression, $\varepsilon_0, ..., \varepsilon_n$ les termes d'erreur et n le nombre d'observations. Ainsi, dans le modèle classique, les coefficients β_k sont considérés identiques dans toute la zone d'étude, ce qui peut occulter la richesse géographique d'un phé-

nomène qui varie dans l'espace. À l'inverse, la RGP possède des coefficients variables qui dépendent des coordonnées géographiques des observations. Ainsi, les β_k sont des surfaces continues qui seront estimées en fonction de certains points de l'espace étudié :

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) x_{ik} + \varepsilon_i,$$

où (u_i, v_i) sont des coordonnées géographiques.

2.4.1 Estimation et paramètres du modèle

Le principe de la RGP est de diminuer l'importance des observations les plus éloignées en accordant à chaque observation un poids décroissant selon sa distance avec le point d'intérêt. Ainsi, comme les données sont géographiquement pondérées, les observations les plus proches exercent une influence plus importante dans l'estimation des coefficients de régression locaux, par rapport aux observations les moins proches. Pour tous les points i de coordonnées (u_i, v_i), chaque ensemble de coefficients de régression est estimé par la méthode des moindres carrés pondérés. Ainsi, l'expression matricielle de l'estimation des coefficients est définie par :

$$\hat{\beta}_{(u_i,v_i)} = (X^\top W_{(u_i,v_i)} X)^{-1} X^\top W_{(u_i,v_i)} Y,$$

où $\mathbf{X}$ est la matrice des variables explicatives (avec une colonne de 1 pour les constantes), $\mathbf{Y}$ est le vecteur de la variable dépendante, $\hat{\beta} = (\beta_{i0}, ..., \beta_{im})^\top$ est le vecteur des $m+1$ coefficients locaux de régressions et $W(u_i, v_i)$ est la matrice diagonale indiquant la pondération géographique de chaque donnée observée pour le point i de coordonnées (u_i, v_i).

Par rapport aux observations plus lointaines, les observations proches du point i sont supposées exercer davantage d'influence sur les paramètres estimés à l'endroit i. Dans la matrice $W(u_i, v_i)$, le poids des observations est donc décroissant avec la distance au point i. Une fonction noyau détermine cette décroissance du poids de chaque observation avec la distance au point d'origine. Il existe trois paramètres clés pour spécifier la fonction de noyau adéquate : la forme du noyau, la taille de la bande passante et le choix d'un noyau fixe ou adaptatif.

Gaussien	$w_{ij} = \exp(-\frac{1}{2}(\frac{d_{ij}}{b})^2)$.
Exponentiel	$w_{ij} = \exp(-\frac{\lvert d_{ij}\rvert}{b})$.
Box-car	$w_{ij} = \begin{cases} 1 & \text{si } \lvert d_{ij}\rvert < b, \\ 0 & \text{sinon.} \end{cases}$
Bicarré	$w_{ij} = \begin{cases} (1-(d_{ij}/b)^2)^2, & \text{si } \lvert d_{ij}\rvert < b, \\ 0 & \text{sinon.} \end{cases}$
Tri-cube	$w_{ij} = \begin{cases} (1-(\lvert d_{ij}\rvert/b)^3)^3, & \text{si } \lvert d_{ij}\rvert < b, \\ 0 & \text{sinon.} \end{cases}$

TABLEAU 2.2 – Fonctions noyau où w_{ij} est le $j^{\text{ème}}$ élément de la diagonale de la matrice de poids $W(u_i, v_i)$, d_{ij} la distance entre le point i et le point j, et b est la bande passante. Ces fonctions noyau sont disponibles dans le paquetage *GWmodel*.

La forme du noyau

Si la forme du noyau ne modifie pas beaucoup les résultats (Brunsdon et collab., 1998), il est cependant important de comprendre leurs différences. Il existe des noyaux continus comme les noyaux uniformes, gaussiens ou encore exponentiels qui accordent un poids à toutes les observations. À l'inverse, les noyaux dits à support compact (exp : noyaux Box-Car, Bicarré, Tri-cube) accordent un poids nul aux observations situées au-delà d'une certaine distance. Il est à noter que lorsque le noyau est uniforme, la RGP est identique à une régression par moindres carrés ordinaires (MCO) en chaque point dans la mesure où les poids accordés aux observations sont tous identiques. Afin d'optimiser le temps de calcul qui peut être long lorsque les observations sont nombreuses, le noyau Bicarré est à privilégier. Le tableau 2.2 présente les différentes fonctions noyaux disponibles dans le paquetage *GWmodel* (Gollini et collab., 2015).

Choix d'un noyau fixe ou adaptatif

Le noyau fixe est identique dans toute la zone étudiée et son étendue est déterminée par la distance à partir du point d'intérêt. À l'inverse, l'étendue d'un noyau adaptatif varie selon le nombre de voisins du point d'intérêt. Ainsi, plus la densité des observations est élevée à un endroit, moins le noyau sera étendu.

On privilégiera un noyau fixe avec une répartition uniforme des données et un noyau

adaptatif lorsque celle-ci est non homogène. En effet, lorsque la densité des observations est faible, l'utilisation d'un noyau fixe trop petit n'inclura pas assez d'observations. Si la densité est très élevée, au contraire, un noyau trop grand ne permettra pas de déceler des variations à petite échelle.

La bande passante

La bande passante *h* est le paramètre qui a le plus d'influence sur les résultats, il est donc nécessaire de définir une valeur adaptée aux spécificités des données étudiées. La fonction noyau accorde un poids nul aux observations situées au-delà de la bande passante. Ainsi, plus la valeur de la bande passante est grande, plus le nombre d'observations ayant un poids non nul sera élevé (Floch et collab., 2018). À l'instar du noyau uniforme, lorsque la bande passante tend vers l'infini –et donc inclut l'ensemble des observations, les résultats de la RGP seront similaires à ceux d'une régression par moindres carrés ordinaires.

Il est possible de choisir la valeur de la bande passante en minimisant le critère de validation croisée, qui maximise le pouvoir prédictif du modèle, ou bien en minimisant le critère d'information d'Akaike (AIC) qui favorise un compromis entre la prédiction et la complexité du modèle. Il est important de noter que la valeur de la bande passante qui minimise ces critères est un bon indicateur concernant la pertinence de la RGP sur les données étudiées. En effet, si en minimisant ces critères la valeur optimale de la bande passante tend vers sa valeur maximale (l'étendue de la zone d'étude ou bien le nombre total d'observations), cela signifie que l'hétérogénéité locale n'est pas significative. Dans ce cas, il n'est pas pertinent d'utiliser une RGP qui a pour but d'appréhender des variations locales de la zone d'étude. À l'inverse, comme présenté par Gollini et collab. (2015), une valeur de bande passante très faible peut indiquer un processus sous-jacent aléatoire.

À titre d'exemple, le tableau 2.4 présente les coefficients d'une régression linéaire classique ainsi que les statistiques descriptives des coefficients d'une RGP pour le jeu de données *EWHP* disponible dans le paquetage *GWmodel*. La variable dépendante *PurPrice*

correspond au prix d'achat d'une propriété en Angleterre ou au pays de Galles en 1999. Les neuf variables explicatives sont présentées dans le tableau 2.3.

On remarque alors que certaines variables ont une influence très différente sur le prix de vente selon la localisation du bien. Par exemple, le coefficient de la variable *TypeFlat*, peut être à la fois négatif avec un minimum de $-28\ 030{,}88$£ et un maximum de 16 024,00£. En d'autres termes, cela signifie que pour certaines localisations, un bien de la catégorie « appartement » aura une valeur estimée plus faible qu'un bien n'étant pas un appartement (ceteris paribus), alors que dans d'autres endroits l'estimation sera plus élevée si le bien est effectivement un appartement. Concernant la variable *FlrArea*, qui correspond à la superficie du bien (m^2), les coefficients sont toujours positifs mais l'impact sur le prix d'un mètre carré supplémentaire n'est évidemment pas le même partout en Angleterre et au Pays de Galles. La carte de la figure 2.13(a) représente le coefficient associé à la variable *FlrArea* au niveau des points où les transactions ont été effectuées. La figure 2.13(b) présente l'estimation de ce coefficient de façon continue sur la carte.

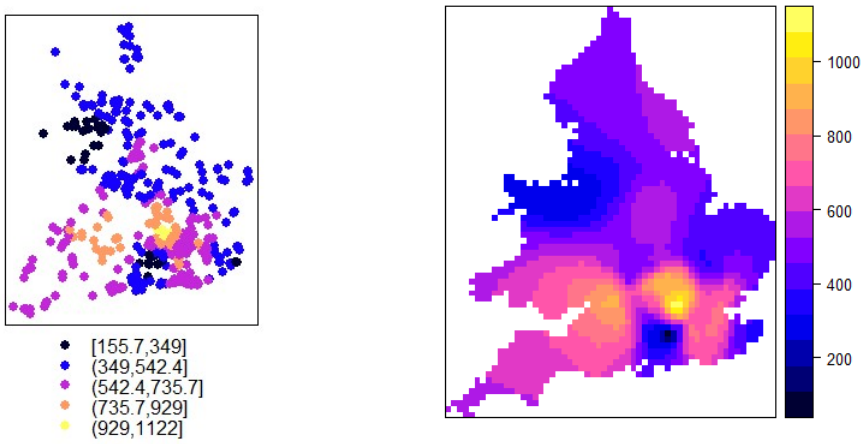
Variables	Description
PurPrice	prix d'achat du bien
BldIntWr	1 si le bien a été construit pendant la guerre mondiale, 0 sinon
BldPostW	1 si le bien a été construit après la guerre mondiale, 0 sinon
Bld60s	1 si le bien a été construit entre 1960 et 1969, 0 sinon
Bld70s	1 si le bien a été construit entre 1970 et 1979, 0 sinon
Bld80s	1 si le bien a été construit entre 1980 et 1989, 0 sinon
TypDetch	1 s'il s'agit d'une maison indépendante, 0 sinon
TypSemiD	1 si le bien est semi-indépendant, 0 sinon
TypFlat	1 si c'est un appartement, 0 sinon
FlrArea	superficie du bien en mètres carrés

TABLEAU 2.3 – Descriptions des variables du jeu de données *EWHP* présent dans le paquetage *GWmodel*.

Pour conclure, l'analyse des données ponctuelles est un sujet très riche avec de multiples applications. Le lecteur pourra se référer aux travaux de Baddeley et collab. (2015) pour approfondir le sujet, notamment sur les modèles de Cox et de Gibbs ou encore sur l'analyse des données situées sur un réseau. L'objectif de ce chapitre était de présenter

	Reg. linéaire		RGP				
Coefficients	Estimation	Valeur-p	Min	1er Q	Med	3ème Q	Max
(Intercept)	10841,44	0,02	−20772,94	−3079,28	4797,80	19443,67	46159,40
BldIntWr	7377,92	0,06	−12745,00	−1852,54	2070,44	7139,44	36096,10
BldPostW	4448,99	0,34	−23093,69	−6147,87	1177,05	5705,65	20900,10
Bld60s	1948,87	0,66	−38345,77	−20827,78	3099,75	7972,36	18361,70
Bld70s	2503,68	0,59	−23904,54	−8084,51	2014,56	8688,09	26139,20
Bld80S	6239,91	0,11	−31889,75	−5176,93	8940,01	14237,21	32846,20
Typdetch	12702,10	0,01	−8629,59	−12907,31	24878,17	33402,50	72829,80
TypSemiD	−12716,37	<0,01	−27778,06	−9879,55	−2475,67	3213,24	28695,00
TypFlat	−15038,31	<0,001	−38030,88	−13226,65	−7810,52	−3867,77	16024,00
FlrArea	585,13	<0,001	155,67	461,76	569,74	700,51	1122,40
	R^2 adj = 0,49		R^2 adj = 0,74				

TABLEAU 2.4 – Résultats de la régression linéaire et de la régression géographiquement pondérée (avec noyau adaptatif Bicarré et bande passante optimisée selon le AIC) du jeu de données *EWHP* présent dans le paquetage *GWmodel*.



(a) Estimations aux observations

(b) Estimation continue

FIGURE 2.13 – Visualisation des estimations des coefficients de la variable *FlrArea* de la RPG (avec noyau adaptatif Bicarré et bande passante optimisée selon le AIC) du jeu de données *EWHP* présent dans le paquetage *GWmodel*.

des concepts fondamentaux liés à l'étude des configurations de points, et de mettre en avant des méthodes pouvant nous aider à étudier l'allocation des dons de la CRC suite aux inondations de la province du Québec en 2019 (application au chapitre 4).

Chapitre 3

Analyse des données surfaciques

L'objectif de ce chapitre est d'introduire les notions fondamentales liées à l'analyse des données surfaciques et de présenter des méthodes pouvant par la suite être appliquées aux données d'inondations récoltées par la CRC (chapitre 4). Dans le cadre de ce type de données, la localisation des observations est considérée comme fixe dans l'espace, et on s'intéresse plutôt à leurs valeurs. Généralement, ces observations sont regroupées selon un ensemble de zones comme le montre la figure 3.1, où chaque unité spatiale représente un secteur de recensement pour la ville de Syracuse dans l'état de New-York, aux États-Unis. Ce chapitre s'appuie partiellement sur le jeu de données *NY_data* présenté par Bivand et collab. (2013) et disponible dans le paquetage *spdep* ainsi que sur les travaux de Le Gallo (2000, 2002), Bivand et collab. (2013), Bouayad Agha et De Bellefon (2018), De Bellefon et collab. (2018) et Floch et Le Saout (2018). Au travers des différentes sections, ce chapitre présentera comment définir une structure de voisinage, puis mettra en avant certaines techniques de détection et de modélisation de l'autocorrélation spatiale au sein des données surfaciques. De même que pour l'analyse des données ponctuelles, ce chapitre n'a pas pour objectif de couvrir toutes les théories et méthodes disponibles, mais plutôt de présenter les concepts fondamentaux et pertinents dans le contexte de l'étude de cas avec la CRC.

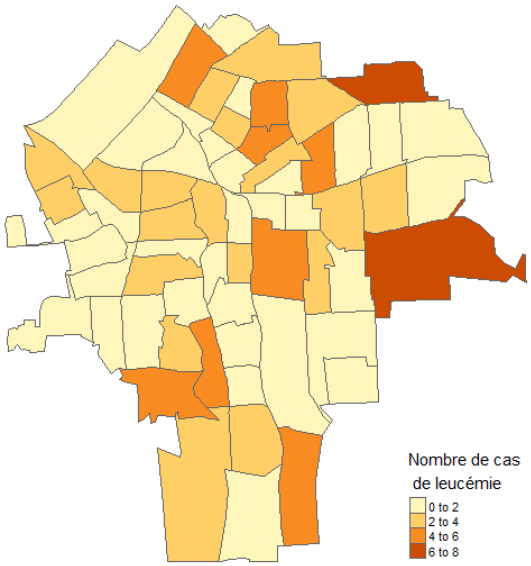


FIGURE 3.1 – Visualisation du nombre de cas de leucémie par secteur de recensement pour la ville de Syracuse (sous-ensemble du jeu de données *NY_data*).

3.1 Définition d'une structure de voisinage

La définition d'une structure de voisinage est une étape essentielle dans la modélisation et l'étude de l'autocorrélation spatiale. Les relations spatiales entres les observations peuvent être définies selon un graphe de voisinage qui pourra ensuite être converti en matrice de voisinage. Cette matrice, de dimension $n \times n$, où n est le nombre d'observations, évalue la similarité entre les observations et, dans le cas d'une matrice de poids binaire, se définie par :

$$w_{ij} = \begin{cases} 1 & \text{si } i \text{ et } j \text{ sont voisins,} \\ 0 & \text{sinon.} \end{cases}$$

Cette section aura donc pour but de décrire les différentes méthodes pour définir une structure de voisinage et accorder des poids aux éléments voisins.

3.1.1 Définition basée sur la contiguïté

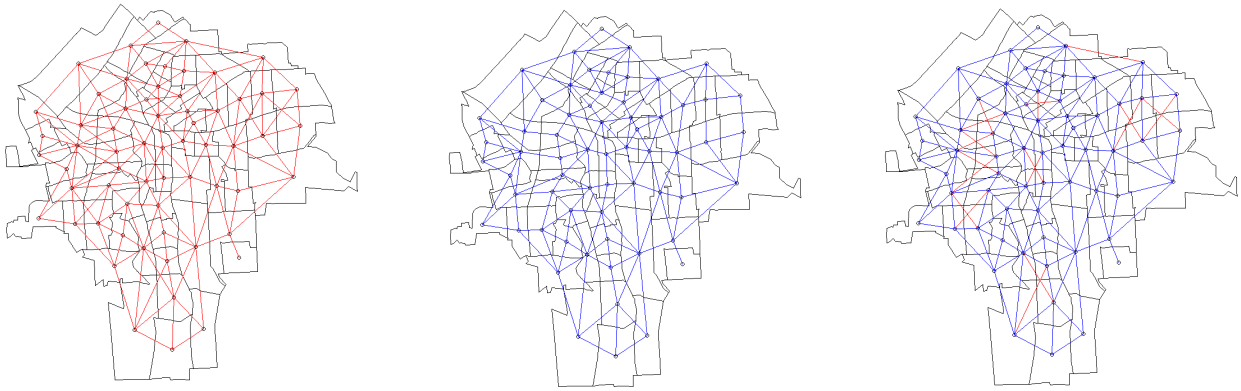
Il est possible de définir une structure de voisinage selon deux types de contiguïté différents : les contiguïtés Queen et Rook. Pour que deux polygones soient considérés comme voisins au sens de la contiguïté Rook, ils doivent posséder au moins un segment de frontière en commun. Dans le cas de la contiguïté Queen, seulement un point en commun permet de les considérer comme des polygones voisins. Pour le jeu de données *Syracuse* (sous ensemble du jeu de donnée *NY_data)* composé de 63 polygones, la contiguïté Queen définit 346 liens de voisinage contre 308 pour la contiguïté Rook (voir figure 3.2(a)). On remarque bien le caractère plus strict de la contiguïté Rook, qui donne lieu à moins de relations de voisinage entre les polygones. Cependant, en présence de nombreux polygones ayant une surface irrégulière, il devient plus complexe d'appréhender les différences entre une contiguïté Queen ou une contiguïté Rook. La figure 3.2(c) illustre cette différence avec en bleu les relations liées par des frontières communes (Rook et Queen) et en rouge les relations de voisinage supplémentaires lorsque deux polygones sont connectés par un même point (uniquement Queen).

Dans certains cas, résumer la proximité entre des polygones par la distance de leurs centroïdes (voir section suivante) peut conduire à une perte d'une partie de la richesse de l'information spatiale. Il peut donc être pertinent de définir une structure de voisinage selon la contiguïté que les polygones observent entre eux. Ainsi, dans le cadre d'études sur des données démographiques et sociales, la structure de voisinage la plus pertinente est souvent définie selon la contiguïté, dans la mesure où la séparation par une frontière administrative peut avoir plus d'importance qu'une mesure de distance.

3.1.2 Définition basée sur la distance

Définition basée sur des graphes (géométriques)

La triangulation de Delaunay est une méthode géométrique permettant de relier les points (par exemple les centroïdes des polygones) sous forme de triangle afin de maximiser l'angle minimal de l'ensemble des triangles. Il existe aussi différents sous-graphes



(a) QUEEN (346 liens) (b) ROOK (308 liens) (c) QUEEN + ROOK

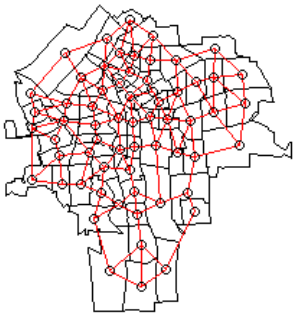
FIGURE 3.2 – Contiguïté QUEEN et ROOK pour la ville de Syracuse de l'état de New-York, États-Unis.

de la triangulation de Delaunay comme celui de la sphère d'influence, de Gabriel et des voisins relatifs.

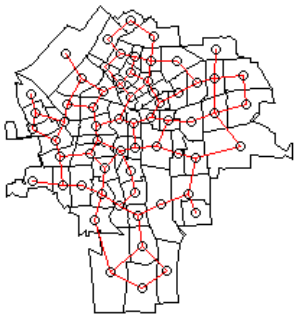
Le graphe de la sphère d'influence relie deux points (ici centroïdes des polygones) si leurs cercles de voisin les plus proches se coupent. En d'autres termes, des points sont voisins si des cercles centrés sur ces points, et avec un rayon égal aux distances de leur plus proche voisin, se croisent (Avis et Horton, 1985). Cette configuration ne garantie pas de liaison entre tous les points de l'ensemble l'étude, qui ne sont pas forcément liés entre eux. Il est à noter que les voisins issus d'une triangulation de Delaunay ou de la sphère d'influence ont une relation symétrique, c'est-à-dire que si i est un voisin de j, alors j est un voisin de i. En revanche, ce n'est pas le cas pour le graphe de Gabriel (Matula et Sokal, 1980) qui relie deux points i et j si et seulement si tous les autres points sont en dehors du cercle de diamètre $[i, j]$. De manière similaire, le graphe des voisins relatifs (Toussaint, 1980) ne garantie pas la symétrie entre les voisins. Ce graphe considère que deux points i et j sont voisins si $d(i, j) \leq \max[d(i,k), d(j,k)]$ pour tout $k = 1, ..., n$ et $k \neq i, j$, avec $d(i, j)$ la distance entre i et j. Le graphe des voisins relatifs assigne moins de liens de voisinages que les autres configurations (voir figure 3.3).

(a) Triangulation de Delaunay (350 liens)

(b) Graphe de la sphère d'influence (294 liens)



(c) Graphe de Gabriel (131 liens)



(d) Graphe des voisins relatifs (83 liens)

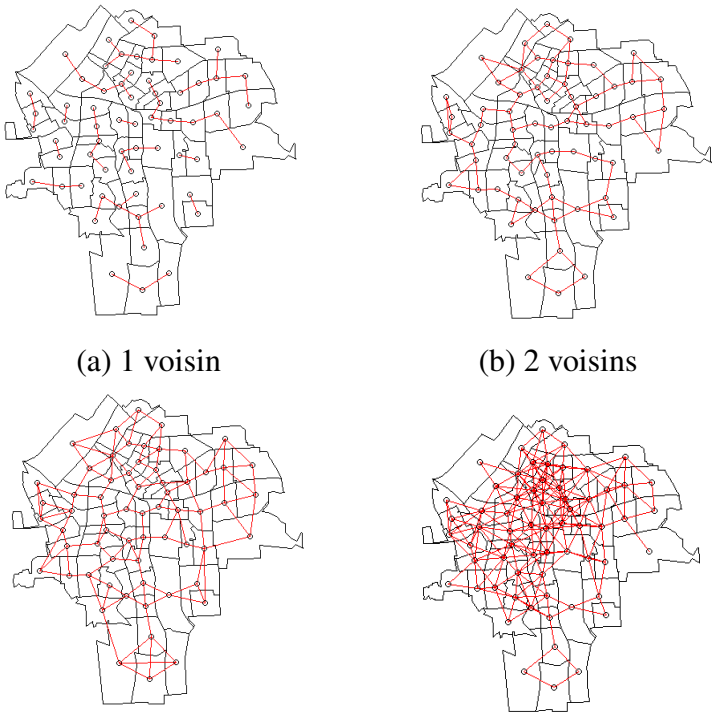
FIGURE 3.3 – Graphes de voisinage fondés sur des notions géométriques pour la ville de Syracuse de l'état de New-York, États-Unis.

Définition basée sur les plus proches voisins

Les graphes de voisinage basés sur la notion des voisins les plus proches permet de sélectionner comme voisins les k voisins les plus proches. Ainsi, tous les éléments ont au moins un voisin, ce qui est généralement plus représentatif de la réalité car il est rare de trouver des zones géographiques totalement isolées. Cependant, il n'est pas aisé de déterminer la valeur de k qui reflète le mieux les relations spatiales sous-jacentes. De plus, à l'instar des graphes de Gabriel et de celui des voisins relatifs, les graphes basés sur les plus proches voisins sont souvent asymétriques.

Il est aussi possible de sélectionner uniquement les points situés à une certaine distance. On peut alors utiliser la distance minimale (d_{min}) pour laquelle chaque point a au moins un voisin, et garder comme voisins les points situés entre 0 et d_{min}. Cependant, la méthode de la « distance minimale » est peu adaptée aux données avec des zones espacées

car il risque d'y avoir de grandes disparités dans le nombre de voisins (Bivand et collab., 2013). C'est notamment ce que l'on pourra observer pour les données de la CRC (chapitre 4) qui s'étendent sur une grande partie du Québec, avec des observations parfois très éloignées des autres. En effet, il suffit d'un point relativement isolé pour que cette distance d_{min} soit beaucoup plus élevée que la distance du plus proche voisin pour de nombreux points situés dans une zone plus dense.



(a) 1 voisin (b) 2 voisins

(c) 3 voisins (d) Voisins à distance minimale

FIGURE 3.4 – Graphes de voisinage fondés sur les plus proches voisins pour la ville de Syracuse de l'état de New-York, États-Unis.

3.1.3 Autres méthodes

Il existe d'autres méthodes moins courantes et plus spécifiques pour codifier la structure de voisinage. Il est par exemple possible de définir les voisins en se basant sur l'optimisation d'une trajectoire selon le problème du voyageur de commerce ou bien selon

la méthode *general randomized tessellation stratified* (Stevens Jr et Olsen, 2004). Ces méthodes sont notamment populaires dans le cadre d'un échantillonage spatial.

3.1.4 Attribution de poids

Lorsque les relations spatiales entre les observations sont définies, on peut transformer la liaison entre les points i et j en l'élément w_{ij} de la matrice de poids W de dimension $n \times n$. Anselin et Griffith (1988) définissent cette matrice comme l'expression formelle de la dépendance spatiale entres les observations.

Il existe différentes manières de spécifier une matrice de poids. La matrice de poids binaire est la plus courante et se définit, comme nous l'avons vu précédemment, de cette façon :

$$w_{ij} = \begin{cases} 1 & \text{si } i \text{ et } j \text{ sont voisins,} \\ 0 & \text{sinon.} \end{cases}$$

La distance entre les zones peut aussi être prise en compte dans l'élaboration de la matrice de poids, qui ne nécessite donc pas une matrice de voisinage au préalable. En effet, on peut utiliser une distance maximale à partir de laquelle les poids sont fixés à 0 ($w_{ij} = 1$ si d est $< d_0$ et $w_{ij} = 0$ sinon). En établissant un tel critère, cela permet alors de limiter la variabilité dans le nombre de voisins lorsque les tailles des zones sont hétérogènes. Il est aussi possible d'utiliser une fonction de la distance pour attribuer les poids. On peut citer la fonction exponentielle négative ($w_{ij} = e^{-\alpha d}$) ou encore la fonction de l'inverse de la distance ($w_{ij} = d_{ij}^{-\alpha}$ si $d_{ij} < d_0$ et $w_{ij} = 0$ sinon), avec α un paramètre estimé ou préalablement définit.

Pour finir, il est aussi possible de tenir compte de la force des relations entres les zones. On peut par exemple définir le poids par $b_{ij}^{\alpha}/d_{ij}^{\beta}$, où b_{ij} est une mesure de la force de la relation entre les zones i et j. Cette mesure n'est pas nécessairement symétrique et peut par exemple représenter le pourcentage de frontières communes.

Normalisation de la matrice de poids

Afin de ne pas créer une hétérogénéité entre les zones qui aurait un impact sur les résultats des tests d'autocorrélation spatiale, il est important d'effectuer une normalisation de la matrice de poids. Cette hétérogénéité est issue du « degré de liaison » qui représente la somme du poids des voisins d'une zone et qui dépend ainsi de son nombre de voisin. On peut normaliser les poids de différentes manières (Tiefelsdorf et collab., 1999) :

- Avec une **normalisation en ligne**, chaque poids est divisé par la somme des poids des voisins de la même zone. Ainsi, pour chaque zone, la somme des poids $\sum_{j=1}^{n} w_{ij}$ est égale à 1. Par conséquent, cela induit une certaine compétition entre les voisins, dans la mesure où plus une zone a de voisins, plus leurs poids sont faibles.

- Avec une **normalisation globale**, les poids sont standardisés pour que la somme de l'ensemble des poids soit égale au nombre total de zones. Ainsi, les poids sont multipliés par $\frac{n}{\sum_{j=1}^{n}\sum_{i=1}^{n} w_{ij}}$.

- Avec la **normalisation uniforme**, les poids sont standardisés pour que la somme de l'ensemble des poids soit égale à 1 : $\sum_{j=1}^{n}\sum_{i=1}^{n} w_{ij} = 1$.

- La **normalisation par stabilisation de la variance** est un peu plus complexe que les autres et a été introduite par Tiefelsdorf et collab. (1999) dans le but de réduire l'hétérogénéité dans les poids liée aux différences de tailles et de voisins entre les zones.

Il est à noter que lorsque l'on ne connaît pas vraiment le processus spatial supposé, il peut être préférable de ne pas trop s'éloigner d'une matrice de poids binaire (Bavaud, 1998).

Suite à la spécification d'une matrice de poids, on peut finalement obtenir une variable spatialement décalée Wx ou Wy, qui correspond à la moyenne pondérée par W des valeurs des observations voisines. Ces variables jouent un rôle déterminent dans la modélisation de l'autocorrélation spatiale, détaillée dans les sections suivantes.

3.2 Autocorrélation spatiale

Avec ce qui est souvent qualifié de première loi de la géographie, Tobler (1970) affirme que « tout interagit avec tout, mais deux objets proches ont plus de chance de le faire que deux objets éloignés ». C'est ainsi que très fréquemment, les valeurs d'observations géolocalisées se caractérisent par des dépendances spatiales qui sont de plus en plus intenses lorsque les localisations sont de plus en plus proches. Après avoir décrit l'autocorrélation spatiale et ses enjeux, cette section abordera les propriétés et méthodes liées à la détection des deux sortes d'autocorrélation spatiale : globale et locale.

3.2.1 Définition de l'autocorrélation spatiale

L'autocorrélation correspond à la corrélation (positive ou négative) d'une variable avec elle-même, lorsque les observations sont décalées dans l'espace (autocorrélation spatiale) ou dans le temps (autocorrélation temporelle). L'autocorrélation spatiale est ainsi définie comme :

- **Positive** si les valeurs des lieux proches se ressemblent davantage que celles des lieux éloignés. On observe alors le regroupement d'observations semblables.
- **Négative** si les valeurs des lieux proches sont plus différentes que celles des lieux éloignés. On observe alors le regroupement d'observations dissemblables.
- **Non existante** si aucun lien n'existe entre la valeur des observations et leur proximité dans espace. On observe alors une répartition spatiale aléatoire des observations.

En statistique, de nombreux tests et analyses reposent sur l'hypothèse d'indépendance des variables. La présence d'autocorrélation, entraîne donc l'abandon de cette hypothèse fondamentale d'indépendance des observations. De plus, l'autocorrélation spatiale diffère de l'autocorrélation temporelle dans la mesure où cette dernière est unidirectionnelle

(seul le passé influence le futur) alors que l'autocorrélation spatiale est multidirectionnelle. Cet aspect implique une complexité beaucoup plus élevée, notamment concernant l'interprétation des résultats.

Pour finir, on identifie deux origines principales de l'autocorrélation spatiale. En premier, elle peut résulter de processus non observés qui associent des lieux différents ensemble et qui sont ainsi à l'origine d'une organisation particulière des activités dans l'espace (Le Gallo, 2000). Dans un second temps, la détection d'une autocorrélation spatiale peut signifier une mauvaise spécification du modèle, avec par exemple une omission de certaines variables spatialement corrélées, un mauvais choix d'échelle ou encore des données manquantes. L'autocorrélation peut alors servir comme un outil de détection d'une spécification inadéquate d'un modèle.

3.2.2 Dépendance spatiale globale

Le contenu de cette section a pour but de présenter les principaux outils de détection d'une dépendance spatiale globale.

Le diagramme de Moran

Le diagramme de Moran est un premier outil intéressant pour comprendre une structure spatiale. Il s'agit d'un graphique avec, pour abscisse, les valeurs de la variable d'intérêt Y, et pour ordonnée, la variable spatialement décalée Wy. Elle correspond au produit entre les valeurs de cette variable pour les observations voisines et la matrice de poids normalisée W. La droite de régression de Wy en fonction de y est aussi représentée et sa pente correspond au paramètre ρ qui définit l'autocorrélation spatiale.

Divisé par quadrants, ce diagramme permet de situer et d'identifier une structure spatiale selon la prépondérance des observations dans l'un des quatre quadrants. Il permet aussi d'identifier s'il y a des observations qui s'éloignent beaucoup de la structure spatiale dominante. Il existe ainsi quatre structures spatiales différentes :

- **Structure High-High** (quadrant en haut à droite) : lorsque les observations y ont une valeur plus élevée que celle de la moyenne de leurs voisins (Wy), dont la valeur est relativement similaire. L'autocorrélation spatiale est positive avec un indice élevé.

- **Structure High-Low** (quadrant en bas à droite) : lorsque les observations y ont une valeur plus élevée que celle de la moyenne de leurs voisins (Wy), dont la valeur est très différente. L'autocorrélation spatiale est négative avec un indice élevé.

- **Structure Low-Low** (quadrant en bas à gauche) : lorsque les observations y ont une valeur plus faible que celle de la moyenne de leurs voisins (Wy), mais dont la valeur est relativement similaire. L'autocorrélation spatiale est positive avec un indice faible.

- **Structure Low-High** (quadrant en haut à gauche) : lorsque les observations y ont une valeur plus faible que celle de la moyenne de leurs voisins (Wy), dont la valeur est très différente. L'autocorrélation spatiale est négative avec un indice faible.

Si une structure dominante est identifiée, on peut ensuite quantifier cette autocorrélation spatiale et étudier sa significativité avec des indices de corrélation. Ces indices permettent de spécifier la corrélation entre les valeurs géographiquement voisines d'une variable étudiée.

La figure 3.5 illustre ce diagramme pour le jeu de données *NY_data*, dont *Syracure* était un sous ensemble, avec comme variable d'intérêt le nombre de cas de leucémie par secteur de recensement dans l'état de New-York, États-Unis. La pente de la droite de régression est positive mais une structure dominante est difficilement identifiable.

L'indice de Moran

L'indice de Moran, aussi appelé I de Moran (*Moran's I*) se calcule de cette manière :

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i-\overline{y})(y_j-\overline{y})}{\sum_{i=1}^{n}(y_i-\overline{y})^2},$$

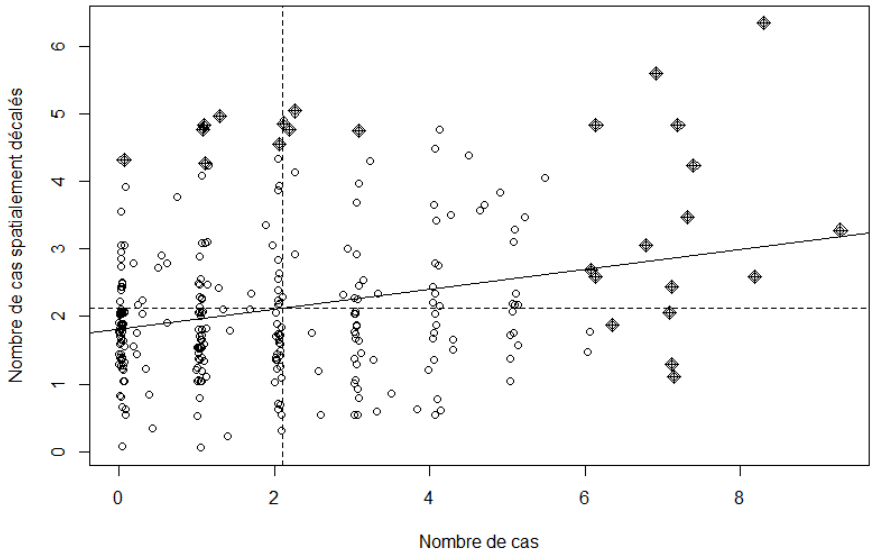


FIGURE 3.5 – Diagramme de Moran pour la distribution des cas de leucémie dans l'état de New-York, États-Unis (jeu de données *NY_data*).

où y_i est la i$^{\text{ème}}$ observation, $\bar{y}$ est la moyenne de la variable d'intérêt et w_{ij} le poids attribué au lien entre i et j.

La valeur de cet indice n'est autre que celle de la pente de la droite du diagramme de Moran. L'autocorrélation spatiale est ainsi positive lorsque l'indice est supérieur à zéro et négative lorsqu'il y est inférieur. Cependant, aurait-il pu être possible, grâce au hasard, d'obtenir des valeurs aussi similaires (au dissemblables) pour les observations voisines ? Les tests d'hypothèses permettent d'étudier la significativité de ce résultat.

Test d'hypothèses

On peut alors tester l'absence d'autocorrélation spatiale pour la variable étudiée :

H_0 : absence d'autocorrélation spatiale,

H_1 : présence d'autocorrélation spatiale.

Pour effectuer ce test, il est nécessaire de préciser la distribution de la variable d'intérêt sous H_0, qui peut se baser sur une hypothèse de normalité ou de randomnisation (Monte Carlo). Pour la première, chacune des valeurs de la variable est le résultat d'un

tirage indépendant dans la distribution normale propre à chaque zone géographique. Pour la seconde, qui est l'hypothèse la plus fréquemment utilisée, le I de Moran est calculé sur une nouvelle version du jeu de données où les valeurs sont assignées de manière aléatoire (par permutations) aux zones géographiques. Cette opération est répétée plusieurs fois afin d'établir une distribution des valeurs attendues sous l'hypothèse nulle H_0. Par la suite, la valeur observée du I de Moran est comparée à cette distribution simulée pour étudier si les valeurs observées de la variable peuvent être considérées comme aléatoires. Sous l'hypothèse H_0, la statistique de test de Wald $\frac{I-\mathrm{E}(I)}{\sqrt{\mathrm{V}ar(I)}}$ suit asymptotiquement une loi normale $\mathscr{N}(0,1)$.

L'indice de Geary (Geary, 1954) permet aussi de quantifier l'autocorrélation spatiale en étudiant le rapport entre la variance des régions voisines et la variance totale, mais il est généralement moins stable que le I de Moran (Upton et collab., 1985). Il est définit par :

$$C = \frac{n-1}{2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - y_j)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \quad i \neq j,$$

avec n le nombre d'unité spatiale, y_i est la i$^{\text{ème}}$ observation, $\overline{y}$ est la moyenne de la variable d'intérêt et w_{ij} le poids attribué au lien entre i et j. Cet indice s'étend de 0 à 2. Il indique une autocorrélation spatiale positive s'il est inférieur à 1 et négative s'il est supérieur à 1.

La définition de la structure du voisinage (section 3.1) a un impact important dans la mesure de l'autocorrélation spatiale car ces indices comparent la valeur d'une variable pour une zone avec la valeur de ses voisins. Ainsi, plus la structure de voisinage englobe un nombre important de voisins, plus la moyenne des valeurs des voisins se rapprochera de la valeur moyenne dans tout le jeu de données étudié.

Concernant le jeu de donnée *NY_data*, on obtient un indice de Moran de 0,15 et l'hypothèse nulle (absence d'autocorrélation spatiale) est rejetée.

3.2.3 Dépendance spatiale locale

Les statistiques d'autocorrélation globales font l'hypothèse de la stationnarité du processus spatial et supposent ainsi que l'autocorrélation serait la même dans toute la surface étudiée. Cependant, cette hypothèse est généralement peu réaliste, notamment lorsque le nombre d'observations est élevé. Ainsi, ces indicateurs globaux nous informent uniquement sur le processus global et n'indiquent ni la localisation des regroupements, ni leur nature (*hotspot, coldspot, valeurs extrêmes*). De plus, l'absence d'autocorrélation globale ne signifie pas forcément une absence de regroupements. En effet, leurs signes peuvent s'annuler mutuellement, ce qui amènerait à une autocorrélation faible ou nulle.

Dans cette section, nous allons étudier comment les indicateurs locaux, appelés Indicateurs d'Autocorrélation Spatiale Locale (LISA), permettent de décomposer les indicateurs globaux afin de connaître la contribution de chaque observation à l'autocorrélation spatiale globale.

Ces indicateurs, définis par Anselin (1995) ont deux objectifs. D'une part, ils peuvent détecter les regroupements significatifs de valeurs identiques autour d'une localisation particulière. D'autre part, ils peuvent repérer les zones de non-stationnarité spatiale, qui ne suivent pas le processus global. Pour qu'un indicateur soit considéré comme un LISA, il doit pouvoir indiquer, pour chaque observation, l'intensité du regroupement de valeurs semblables (ou opposées). De plus, la somme des indices locaux sur l'ensemble des observations doit être proportionnelle à l'indice global correspondant. Le I de Moran local correspond à ces critères et est généralement le plus utilisé.

Indice de Moran Local

L'indice de Moral local, ou I de Moran local, est définit par :

$$I_i = (y_i - \overline{y}) \sum_{j=1}^{n} w_{ij}(y_j - \overline{y}).$$

On a donc $I_{\text{Global}} = \alpha * \sum_{i=1}^{n} I_i$.

Lorsque I_i est supérieur à 0, il indique un regroupement de valeurs similaires qui sont plus élevées ou plus faibles que la moyenne. À l'inverse, lorsque cet indice est inférieur à 0, alors on est en présence d'un regroupement de valeurs dissemblables avec par exemple, des valeurs faibles parmi des valeurs élevées.

Lorsqu'il n'y a pas d'autocorrélation spatiale globale, les LISA vont identifier les zones où les valeurs qui se ressemblent se regroupent de manière significative. Dans ce cas, les liaisons entre zones voisines sont particulièrement fortes. Dans le cas où une autocorrélation globale est aussi présente, les LISA vont détecter les zones qui influent particulièrement sur le processus global ou qui s'en démarquent. Certaines zones peuvent en effet posséder une structure spatiale opposée à celle du processus global.

Concernant la significativité des résultats, il est nécessaire de faire preuve de prudence quant à l'interprétation des valeurs-p calculées sous l'hypothèse de normalité. En effet, pour chaque I_i, le test de significativité se base sur une statistique supposée suivre asymptotiquement une loi normale sous H_0. Sous l'hypothèse nulle, plusieurs répartitions aléatoires sont simulées puis leurs indicateurs locaux sont calculés afin de tester la validité de l'hypothèse de normalité. Or, Anselin (1995) démontre qu'en présence d'autocorrélation spatiale globale, l'hypothèse de normalité des I_i n'est plus vérifiée. Par ailleurs, plus il y a d'indices d'autocorrélation spatiale locaux, plus le risque d'obtenir un résultat significatif par hasard augmente. En d'autres termes, cela augmente le risque de conclure à une autocorrélation spatiale locale qui n'existe pas en réalité (Anselin, 1995). Il est cependant possible de limiter ce risque en utilisant différentes méthodes pour ajuster les valeurs-p. Certaines entre-elles, comme la méthode de Holm (1979), vont privilégier un risque faible de détection d'une autocorrélation locale, quitte à ne pas détecter un *cluster* local. Ainsi, le choix de la méthode d'ajustement doit dépendre des spécificités et des risques liés l'étude.

3.3 Modélisation

Lorsqu'une autocorrélation spatiale dans les données est détectée, différents modèles économétriques permettent d'en tenir compte dans l'étape de modélisation. À partir d'une même définition de la structure de voisinage, ces modèles sont donc en concurrence et plusieurs règles de décision permettent de déterminer le modèle le plus adéquat pour décrire et interpréter l'autocorrélation spatiale détectée. La comparaison de ces modèles offre aussi un moyen d'analyser leur robustesse relative à l'incertitude du processus ayant généré les données, qui reste inconnu. Cette section a pour but d'introduire les notions liées à la spécification et aux critères de décision des principaux modèles d'économétrie spatiale.

3.3.1 Modèles d'économétrie spatiale

Dans cette section, nous allons présenter différents modèles d'économétrie spatiale qui, sur le même principe que la régression linéaire classique, visent à évaluer l'association entre une variable Y et une ou plusieurs variable(s) X.

Pour commencer, l'intégration de l'autocorrélation spatiale dans les modèles d'économétrie spatiale peut se faire de différentes manières. En effet, l'autocorrélation spatiale se manifeste via des variables spatialement décalées (endogènes et exogènes) ou bien via les termes d'erreurs. Par rapport au modèle de régression linéaire classique (où les paramètres sont estimés par la méthode des moindres carrés ordinaire) $Y = X\beta + \varepsilon$, les modèles d'économétrie spatiale incorporent d'autres termes permettant de prendre en compte la dépendance spatiale des observations. Cette section décrit en premier les modèles les plus couramment utilisés, puis présente ensuite une généralisation de ces différents modèles d'économétrie spatiale.

- **Modèle autorégressif spatial (*Spatial AutoRegression*, SAR)**
  Dans le modèle SAR, une variable endogène décalée est introduite dans le modèle de régression linéaire classique :

$$Y = \rho Wy + X\beta + \varepsilon,$$

où Wy représente la variable endogène décalée pour la matrice de poids W et ρ le paramètre autorégressif spatial représentant l'intensité de l'interaction entre les observations de y. Ainsi, l'observation y_i de la zone i est partiellement expliquée par les valeurs de y des zones voisines de i : $Wy_i = \sum_{j \neq i} w_{ij} y_j$.

Si une variable endogène décalée est présente dans le processus générateur des données, les estimateurs des MCO dans le modèle de régression linéaire classique seront biaisés et non convergents.

- **Modèle à interactions exogènes (*Spatial Lag X*, SLX)**
  Dans le modèle SLX, on inclut cette fois-ci une (ou plusieurs) variable(s) exogène(s) décalée(s) dans le modèle MCO :

$$Y = X\beta + WX\theta + \varepsilon,$$

où θ est le vecteur des paramètres spatiaux indiquant l'intensité des effets d'interaction exogène et WX l'ensemble des variables exogènes décalées pour la matrice de poids W. Ainsi, l'observation y_i de la région i est partiellement expliquée par les valeurs de X des régions voisines de i.

- **Modèle à erreur autocorrélées spatialement (*Spatial Error model*, SEM)**
  Le modèle SEM se concentre sur l'autocorrélation spatiale des erreurs et se définit par :

$$Y = X\beta + u,$$
$$u = \lambda Wu + \varepsilon,$$

où λ représente l'intensité de l'interdépendance entre les résidus qui correspond à l'effet de corrélation spatiale des erreurs. Si, à tord, on ne prend pas en compte

une autocorrélation spatiale des erreurs, les estimateurs ne seront pas biaisés mais inefficients.

La présence d'autocorrélation spatiale des erreurs peut être le signe d'une mauvaise spécification du modèle où par exemple des variables pertinentes sont omises. Les erreurs captent alors l'effet qui n'est pas décelé par les variables explicatives présentes dans le modèle. Ainsi, la présence d'autocorrélation spatiale peut indiquer l'existence de variables significatives qui, dans l'idéal, devraient être inclues dans le modèle.

- **Modèle spatial de Durbin (*Spatial Durbin model*, SDM)**
  À l'inverse des modèles SAR, SLX et SEM qui incluent chacun un seul type de dépendance spatiale, le modèle SDM incorpore à la fois des interactions endogènes et exogènes :

$$Y = \rho W y + X\beta + WX\theta + \varepsilon.$$

Finalement, les modèles présentés ci-dessus peuvent se généraliser sous la forme suivante (à partir du modèle fondateur de Manski (1993)) :

$$Y = \rho WY + X\beta + WX\theta + u,$$

$$u = \lambda Wu + \varepsilon.$$

Ainsi, selon les contraintes utilisées ($\rho = 0$, $\theta = 0$, $\lambda = 0$), on retrouve les modèles SAR, SLX, SEM et SDM. D'autres modèles comme le *Spatial Durbin Error Model* (SDEM) ou encore le *Spatial Autoregressive Consused* (SAC) (Kelejian et Prucha, 2010), contraints respectivement sous $\rho = 0$ et $\theta = 0$, peuvent être déduits de ce modèle mais sont moins courants.

3.3.2 Tests et critères statistiques pour le choix du modèle

Lorsque l'on ne connaît pas au préalable les types d'interactions spatiales présents dans les données (la forme prise par l'autocorrélation spatiale), des tests statistiques per-

mettent d'orienter le choix vers un modèle plutôt qu'un autre. Avant de présenter les différentes approches de sélection de modèles, voici les tests statistiques utilisés :

Test du rapport de vraisemblance (*Likelihood-Ratio*) : Il s'agit du double de la différence entre la fonction de log-vraisemblance évaluée sous H_1 (modèle non-contraint) et la fonction de log-vraisemblance évaluée sous H_0 (modèle contraint) qui peut ainsi s'exprimer par $-2[\log\{L(\theta_0)\} - \log\{L(\hat{\theta})\}]$ où θ_0 est le maximum de vraisemblance contraint sous H_0 et $\hat{\theta}$ est le maximum de vraisemblance du modèle sans restriction. La statistique du test suit une loi de χ^2 avec un nombre de degrés de liberté égal au nombre de contraintes imposées sous H_0. Dans le contexte des modèles présentés ci-dessus, on va par exemple pouvoir utiliser le test LR_θ pour comparer le SDM (modèle non-contraint) avec le SAR (modèle contraint avec $\theta = 0$). Aussi, le test de l'hypothèse nulle $H_0 : \theta = -\rho\beta$, appelé test du rapport de vraisemblance de l'hypothèse de facteur commun, peut permettre de comparer un SDM et un SEM en utilisant le rapport de vraisemblance.

Tests du multiplicateur de Lagrange : Ces tests sont particulièrement aisés à mettre en oeuvre dans la mesure où ils ne nécessitent uniquement que l'estimation par maximum de vraisemblance du modèle contraint (généralement le modèle non-spatial de régression linéaire classique) pour calculer leur statistique de test (convergeant asymptotiquement vers une loi de χ^2 à un degré de liberté). Par exemple, pour tester la présence d'une variable endogène décalée (test proposé par Anselin (1988)), on aura : $H_0 : \rho = 0$ (qui correspond à la RLC) et $H_1 : \rho \neq 0$ (qui correspond au SAR). On pourra aussi tester si les erreurs suivent un processus spatial autoregressif avec l'hypothèse nulle $H_0 : \lambda = 0$. Il existe aussi des versions robustes (RLM_λ ou aussi noté RLM_{ERR} et RLM_ρ, aussi noté RLM_{LAG}) à une mauvaise spécification locale, proposées par Anselin et collab. (1996). Il s'agit par exemple d'ajuster le test LM_λ pour que sa distribution asymptotique reste un χ^2 centré, même en présence locale de ρ. De manière similaire, le RLM_ρ est robuste à une présence locale de λ. Ces tests robustes s'effectuent grâce aux résidus du modèle contraint et ils convergent vers une loi du χ^2 à un degré de liberté.

Différentes approches sont disponibles pour prendre en compte les résultats de ces tests. Premièrement, l'approche ascendante (*bottom-up*) permet de statuer entre trois modèles : le non-spatial (MCO), le SAR et le SEM. Le Gallo (2002) fait une synthèse des travaux de Anselin et Rey (1991), Florax et Folmer (1992) et Anselin et Florax (1995) relatifs aux critères de décision du choix du modèle et propose une démarche générale. Si la valeur-p du test LM_ρ est supérieure à celle du LM_λ et que le RLM_ρ est aussi significatif mais que le RLM_λ ne l'est pas, alors on inclut la variable endogène décalée et le modèle SAR est privilégié ($Y = \rho Wy + X\beta + \varepsilon$). À l'inverse, si la significativité de LM_λ est supérieure à celle de LM_ρ et RLM_λ est significatif mais pas RLM_ρ, alors on inclut une autocorrélation des erreurs et le modèle SEM est sélectionné ($Y = X\beta + u$, avec $u = \lambda Wu + \varepsilon$). Dans le cas où aucun des tests du multiplicateur de Lagrange ne serait significatif, alors le choix devrait se tourner vers le modèle de régression linéaire classique. Si le vrai modèle est un modèle SAR ou SEM, cette approche est la plus pertinente et la plus efficace selon Florax et collab. (2003).

Une seconde approche, proposée par LeSage et Pace (2009), permet de choisir entre le modèle non-spatial, le SAR, le SLX et le SEM. À l'inverse de l'approche ascendante, celle-ci propose de se baser en premier sur le modèle spatial de Durbin (SDM) et de tester la significativité des différents paramètres.

Pour finir, Elhorst (2010) propose une version mixte qui, comme l'approche ascendante, commence avec le modèle non-spatial. Si les tests révèlent la présence d'interactions spatiales ($\rho = 0, \lambda = 0$) alors le modèle spatial de Durbin (SDM) est ensuite estimé. On utilise ensuite les tests du rapport de vraisemblance ($\theta = 0$ ou $\rho = 0$) et/ou du rapport de vraisemblance de l'hypothèse du facteur commun ($\theta = -\rho\beta$) pour choisir entre le modèle SDM ou les modèles SAR, SLX, SEM et le modèle MCO (non-spatial). Dans l'éventualité d'une incertitude, c'est le modèle de Durbin (SDM) qui est privilégié en raison de sa robustesse. Si les observations sont nombreuses, le calcul de la vraisemblance des modèles peut s'avérer assez lourd et LeSage et Pace (2009) détaillent ces enjeux computationnels.

3.3.3 Interprétation des résultats

L'interprétation des résultats diffère selon la nature et les interactions du modèle. En effet, avec un modèle présentant une autocorrélation des erreurs (SEM), l'interprétation des paramètres β des variables explicatives reste identique à celle d'un modèle de régression classique. En revanche, lorsqu'on inclut des variables (endogènes ou exogènes) spatialement décalées, les interprétations sont moins aisées. Dans le cas d'un modèle à interactions exogènes (SLX), les paramètres des variables explicatives permettent de mesurer un effet local : la variation de la valeur d'une variable X dans une zone i va affecter la valeur de y_i, mais va aussi impacter celle de ses voisins. Cependant, une variation n'impactera pas la valeur des voisins de ses voisins, alors que c'est le cas pour un modèle autorégréssif (avec variable spatialement décalée WY). En effet, en présence d'un SAR, le changement d'une variable explicative pour une zone i va directement impacter la valeur de y_i, ce qui va indirectement changer les résultats de toutes les autres zones. On parle alors d'un effet multiplicateur global dans la mesure où il affecte l'ensemble des observations. Cet effet est complexe car la modification de la valeur de y_i impacte les valeurs de ses voisins, ce qui l'impacte ensuite en retour. Il est donc nécessaire de prendre en compte ce phénomène complexe de rétroaction dans l'interprétation des résultats.

De plus, dans la mesure où l'effet marginal de la modification d'une variable explicative est différent pour chaque zone, on calcule plutôt un effet marginal moyen. LeSage et Pace (2009) proposent ainsi de calculer l'effet direct moyen, l'effet total moyen et l'effet indirect moyen. Ces calculs sont complexes et se basent sur des simulations bayésiennes de Monte-Carlo par chaînes de Markov. L'interprétation de l'effet direct moyen peut s'apparenter à celle des coefficients de régression β pour un modèle de régression classique. Celle de l'effet total moyen peut se faire de deux façons. Il peut s'agir de la moyenne des effets sur une zone i suite à la modification d'une unité de la variable X dans toutes les zones, ou bien à l'inverse, de la moyenne des effets d'un changement d'une unité pour la variable X d'une zone i sur l'ensemble des autres zones. Pour finir, l'effet indirect moyen correspond à la différence entre l'effet direct moyen et l'effet total moyen.

Pour conclure, l'interprétation des résultats peut devenir très complexe, notamment en présence d'un modèle combinant plusieurs effets de dépendance (modèles SDM, SDEM ou encore SAC). Ainsi, si l'analyse des tests statistiques et du contexte global de l'étude ne démontrent pas la pertinence d'utiliser ces modèles, il préférable de se contenter d'un modèle de régression classique facile à interpréter.

Chapitre 4

Application aux données d'inondation de la Croix-Rouge canadienne

Après avoir présenté comment analyser des données ponctuelles et surfaciques dans les chapitres 2 et 3, le présent chapitre vise à appliquer ces méthodes aux données récoltées par la Croix-Rouge canadienne suite aux inondations de 2019 au Québec. Dans le cadre de l'analyse des données surfaciques, nous utiliserons aussi partiellement les données récoltées par la CRC dans certaines zones de l'Ontario. L'objectif de cette étude de cas est d'analyser la présence de facteurs ou de phénomènes ayant un impact significatif sur l'allocation des dons aux foyers sinistrés. Ainsi, à l'aide des données de la CRC et de données externes pertinentes à l'égard du contexte d'étude, nous effectuerons des analyses exploratoires puis nous tenterons de modéliser l'aide financière apportée par la CRC aux victimes des inondations de 2019 au Québec. Pour ce faire, la première partie décrit les données et les étapes de pré-traitement nécessaires aux analyses des sections 4.2 et 4.3, qui traiteront ensuite respectivement des données ponctuelles et des données surfaciques.

4.1 Présentation des données et étapes préliminaires

4.1.1 Données de la Croix-Rouge canadienne (CRC)

En collaboration avec la Croix-Rouge Canadienne, ce projet de recherche se base sur des données récoltées suite aux inondations que la province du Québec a vécu durant le printemps 2019. Parmi plusieurs dizaines de champs disponibles dans la base de données constituée par la CRC suite aux versements d'aides financières, les variables suivantes ont été désignées comme pertinentes et assez fiables pour la suite des analyses. L'unité de référence des variables est un foyer sinistré ayant reçu de l'aide de la part de la CRC.

- **Montant total alloué au foyer ($ CAD)** : Il s'agit du montant total en dollar canadien alloué et versé par la CRC au foyer sinistré. *Note : identifié par « montant reçu » dans la suite de l'analyse.*

- **Localisation du foyer sinistré** : Il s'agit des coordonnées (latitude, longitude) reliées à l'adresse du foyer sinistré.

- **Nombre de personnes dans le foyer** : Il s'agit du nombre de personnes enregistrées au sein du foyer sinistré ou plus généralement du nombre d'habitants du logement. En effet, un foyer n'est pas uniquement un foyer familial, mais peut aussi concerner des personnes vivant en colocation. *Note : identifié par « NB in Casefile » dans la suite de l'analyse.*

- **Identifiant du foyer** : Il s'agit de l'identifiant unique de chaque foyer ayant reçu des aides par la CRC.

- **Nombre de personnes de sexe féminin dans le foyer** : Il s'agit du nombre de personnes de sexe féminin enregistrées au sein du foyer sinistré. *Note : identifié par « NB sexe féminin » dans la suite de l'analyse.*

- **Nombre de personnes de sexe masculin dans le foyer** : Il s'agit du nombre de personnes de sexe masculin enregistrées au sein du foyer sinistré. *Note : identifié par « NB sexe masculin » dans la suite de l'analyse.*

- **Nombre d'enfants entre 0 et 5 ans** : Il s'agit du nombre d'enfants âgés de 0 à 5 ans inclusivement dans le foyer sinistré. *Note : identifié par « NB 0 à 5 ans » dans la suite de l'analyse.*

- **Nombre d'enfants entre 5 et 18 ans** : Il s'agit du nombre d'enfants âgés de plus de 5 ans jusqu'à 18 ans inclusivement dans le foyer sinistré. *Note : identifié par « NB 5 à 18 ans » dans la suite de l'analyse.*

- **Nombre de personnes entre 18 et 65 ans** : Il s'agit du nombre de personnes âgées de plus de 18 ans jusqu'à 65 ans inclusivement dans le foyer sinistré. *Note : identifié par « NB 18 à 65 ans » dans la suite de l'analyse.*

- **Nombre de personnes de plus de 65 ans** : Il s'agit du nombre de personnes âgées de plus de 65 ans dans le foyer sinistré. *Note : identifié par « NB + de 65 ans » dans la suite de l'analyse.*

Il est à noter que dans la structure initiale des données de la CRC, chaque observation représentait une personne à part entière et non un foyer. Afin de pouvoir analyser le montant reçu pour chaque foyer ayant un identifiant unique, une restructuration de la base de données a été nécessaire. L'objectif de cette nouvelle structure est de pouvoir traiter les foyers en tant qu'observations. Ainsi, les variables concernant le sexe et l'âge des personnes présentes dans un foyer ont fait l'objet d'une agrégation afin de pouvoir exploiter ces informations au niveau des foyers. Concernant l'âge, le choix a été fait de séparer les personnes selon quatre catégories d'âge : les nourrissons et jeunes enfants, les enfants et adolescents, les adultes et les personnes âgées. Ces variables peuvent ensuite être couplées avec différents types de données externes permettant d'enrichir et d'approfondir les analyses et la valorisation des données récoltées par la CRC.

Collecte de données en situation d'urgence humanitaire

La collecte de données en situation de crise humanitaire est souvent très complexe car elle n'est pas prioritaire face aux opérations de secours et d'aide aux personnes vulnérables. Il peut donc être important de limiter la collecte aux informations essentielles et critiques pour la mise en place des premiers programmes d'aide d'urgence. La CRC collecte ainsi en priorité les informations personnelles (noms, prénoms, numéros de téléphone, adresses, etc.) qui représentent les données les plus importantes dans un contexte d'urgence. Par exemple, dans le cadre d'un programme d'assistance financière à grande échelle où l'objectif est d'aider le plus rapidement possible le plus grand nombre de personnes affectées, peu d'informations sont nécessaires car un montant unique est alloué par foyer. Une récolte de données trop détaillée pourrait ralentir inutilement le processus de déploiement de ce programme et un arbitrage peut alors être nécessaire entre la quantité de données récoltées et la rapidité de l'aide apportée. Par ailleurs, les opérations de secours sont aussi dépendantes du nombre d'équipes bénévoles disponibles et les organismes comme la CRC doivent ainsi opérer avec des ressources limitées. Ce contexte très particulier explique ainsi pourquoi certains champs non-complets de la base de données n'ont pas étés retenus pour être intégrés à cette analyse.

4.1.2 Données externes

Fichier des limites des aires de diffusion du Recensement de 2016 par Statistique Canada

Statistique Canada, l'organisme national de statistique canadien, met à disposition et en libre accès des fichiers des limites géographiques du Canada ainsi que des fichiers numériques des limites (Statistique Canada, 2016a). Ces fichiers sont disponibles pour de nombreuses régions géographiques (provinces et territoires, division de recensement, région économique, etc.). Selon Statistique Canada (2016b), les aires de diffusion (AD) sont les plus petites régions géographiques normalisées pour lesquelles toutes les données du recensement sont diffusées et comptent environ 400 à 700 habitants. Ainsi, l'accès à

ces fichiers de limites est crucial pour l'analyse surfacique et permet aussi de pouvoir utiliser des données socio-démographiques du Recensement national de 2016.

Du fait de leur nature, ces fichiers sont cependant très volumineux et peuvent parfois être lourds à manipuler, surtout lors de visualisations cartographiques. Ils sont disponibles au téléchargement sous plusieurs formats (*.shp, .gml, .tab*).

Données socio-démographiques du Recensement de 2016 par Statistique Canada

En plus des fichiers des limites, Statistique Canada publie tous les cinq ans un portrait statistique national via le programme du Recensement. Celui-ci permet ainsi d'avoir accès à de nombreuses informations sur divers sujets comme la démographie, les revenus, les logements, l'immigration, etc. Dans le cadre de cette recherche, l'intérêt est de pouvoir accéder à des informations à un niveau le plus précis possible (le moins agrégé) pour pouvoir les associer aux observations de la CRC. On a donc sélectionné certaines variables parmi les milliers disponibles (tableau 4.1.2). On utilise le niveau d'agrégation le plus petit (aires de diffusion, AD), pour que ces données puissent correspondre aux polygones du fichier des limites géographiques.

Dans la mesure où les montants versé par la CRC sont accordés en fonction des pertes financières en lien avec le sinistre, et non en fonction de la situation pré-sinistre, l'utilisation de ces variable a notamment pour but de déterminer si elles ont un impact significatif sur les demandes d'aides effectuées par les foyers sinistrés.

Événements d'inondation du printemps 2019 - Données Québec

Le portail de données ouvertes *Données Québec* met à disposition des données relatives aux inondations du printemps 2019 au Québec. Un des jeux de données particulièrement intéressant dans le cadre de cette recherche rassemble des points géolocalisés situant approximativement le centre ou la position la plus significative des événements d'inondation (Données Québec, 2019). Ces informations ont étés recueillies grâce aux observations terrain des conseillers en sécurité civile ou à des intervenants en mesures

Variables	Description
Population 2016	Nombre d'habitants selon l'aire de diffusion
Âge moyen	Âge moyen de la population selon l'aire de diffusion
Taux d'activité	Taux d'activité de la population selon l'aire de diffusion
Taux d'emploi	Taux d'emploi de la population selon l'aire de diffusion
Taux de chômage	Taux de chômage de la population selon l'aire de diffusion
Valeur med logements	Valeur médiane des logements selon l'aire de diffusion (en $ CAD)
Valeur moy logements	Valeur moyenne des logements selon l'aire de diffusion (en $ CAD)
Revenu med ménages	Revenu total médian des ménages en 2015 selon l'AD (en $ CAD)
Taille moy ménages	Taille moyenne des ménages privés selon l'AD

TABLEAU 4.1 – Liste des variables utilisées provenant du Recensement de 2016 par Statistique Canada.

d'urgence pendant les mois d'avril et mai 2019. La figure 4.1 représente l'ensemble des localisations du fichier où des relevés ont été effectués.

La sévérité, classée en quatre catégories (mineure, modérée, importante, extrême), a été relevée au moins une ou plusieurs fois pour chaque point présent dans le fichier *shapefile*. La section 4.1.3 décrira plus amplement les manipulations nécessaires afin d'associer les informations de sévérité aux données de la CRC.

4.1.3 Pré-traitement pour joindre les données externes aux données de la CRC

Plusieurs manipulations sont nécessaires pour associer les données externes aux données collectées par la CRC. Premièrement, il est impératif d'associer à chaque foyer sinistré l'aire de diffusion correspondante. Si les données fournies par la CRC contiennent les codes postaux des observations, il n'est cependant pas possible de lier un code postal à une aire de diffusion. La solution optimale est alors d'utiliser les coordonnées géographiques de chaque point et de les comparer aux limites géographiques des AD fournies par le fichier *shapefile* de Statistique Canada. La fonction `point.in.poly()` du paquetage *spatialEco* permet d'examiner s'il y a une intersection entre les points représentant les habitations sinistrées et les aires de diffusion. Ainsi, chaque observation (point) des

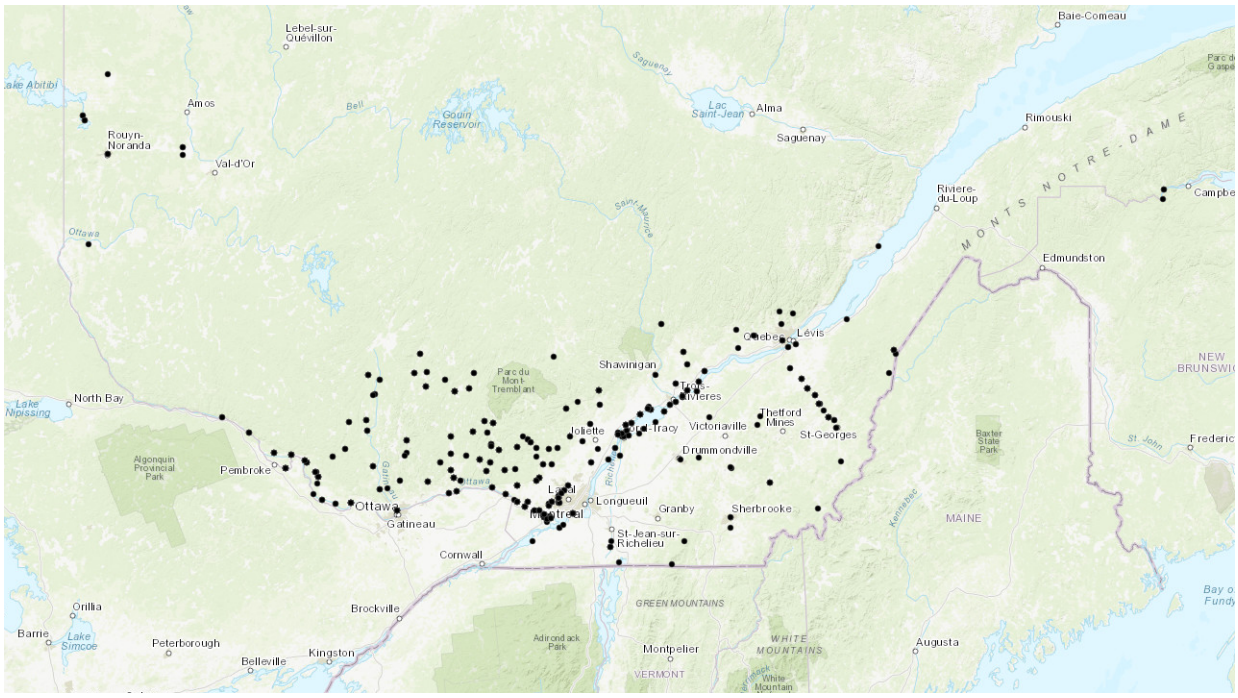


FIGURE 4.1 – Visualisation des positions d'inondations significatives selon *Données Québec*. *Visualisation à l'aide du package* `tmap`.

données de la CRC est associée à une AD. Cela permet ensuite de pouvoir leur attribuer les variables socio-démographiques de Statistiques Canada car leur niveau d'agrégation correspond bien aux aires de diffusion. Il est à noter que Statistique Canada ne publie pas de données de recensement au niveau d'agrégation des codes postaux et qu'il n'existe pas de fichier public des limites géographiques des codes postaux. Ainsi, il est difficilement possible d'effectuer des analyses spatiales à ce niveau géographique au Canada.

Concernant le fichier des événements d'inondation disponibles sur *Données Québec*, l'idée fût d'associer à chaque logement sinistré le niveau de sévérité le plus élevé du point le plus proche. Chaque point représenté sur la figure 4.1 est en fait une superposition d'une ou plusieurs observations relevées à différentes dates. Pour chaque point, on retient uniquement le niveau de sévérité le plus élevé, car même si ce n'est pas celui lié à l'observation la plus à jour, c'est le plus pertinent pour l'évaluation du niveau potentiel des sinistres. Pour adapter cette information au niveau d'agrégation des aires de diffusion, on utilise un principe similaire. Dans la mesure où chaque aire de diffusion sélectionnée

pour l'étude (voir section 4.3.1) ne contient pas toujours une observation localisant un événement d'inondation significatif, nous associons à l'AD le niveau de sévérité du point le plus proche (par rapport à son centroïde).

Enfin, lorsque l'on manipule plusieurs fichiers spatiaux ensemble, il est nécessaire de faire preuve d'une grande précaution quant aux systèmes de coordonnées utilisés (voir section 1.2.3). En effet, des conversions de coordonnées et de système de projection ont été nécessaires car les différents fichiers spatiaux n'étaient initialement pas compatibles. La visualisation en mode « view » du paquetage *tmap* peut s'avérer très utile pour s'assurer de la cohérence des manipulations ou identifier des anomalies.

Traitement des anomalies spatiales

Lors de toute analyse de données, une phase de pré-traitement pour corriger certaines anomalies est presque systématiquement nécessaire. Si la qualité et la précision des coordonnées géographiques des observations de la CRC sont très bonnes, de légères anomalies sont cependant présentes et nécessitent d'être corrigées. En effet, un même couple de coordonnées peut être associé à plusieurs dizaines (jusqu'à 126) foyers différents. Ces localisations correspondent parfois à des lieux publics ou à des boîtes postales. Ainsi, par prudence, au delà de 10 logements situés à un exact même endroit, les observations sont enlevées et ne sont pas prises en compte dans les analyses. Le choix de conserver les observations en dessous ou égal à 10 superpositions permet de ne pas éliminer les adresses où il pourrait y avoir plusieurs logements dans un même bâtiment. Cependant, le choix de 10 habitations est assez arbitraire et s'appuie principalement sur le type de bâtiments fréquemment constaté dans ces régions (peu d'édifices à nombreux étages). On introduit alors de légères perturbations dans les coordonnées de ces points afin qu'ils ne soient plus exactement aux mêmes endroits. Pour de nombreuses méthodes en analyse de données ponctuelles, il est nécessaire d'avoir des coordonnées qui ne se superposent pas.

À titre d'exemple, si un couple de coordonnées est attribué à trois foyers, alors deux de ces foyers recevront des perturbations et le troisième gardera les coordonnées initiales. La figure 4.2 représente la visualisation du jeu de données au complet, c'est à dire des 4

705 foyers, après le traitement de ces anomalies. À cette échelle, les perturbations ne sont pas perceptibles.

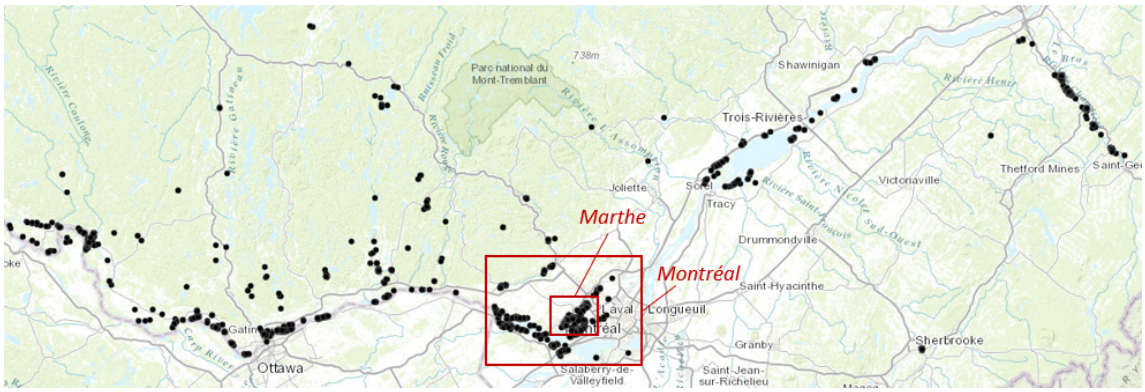


FIGURE 4.2 – Visualisation des foyers ayant reçu une aide de la CRC après le traitement des anomalies.

Traitement des valeurs manquantes

Parmi les données jugées assez fiables (décrites à la section 4.1.1), la variable de l'âge des personnes enregistrées incluait certaines valeurs manquantes. Il a donc été jugé pertinent de supprimer une centaine de foyers (dans tout le jeu de données du Québec) pour lesquels cette information n'était pas disponible.

De plus, pour quelques AD, les variables sélectionnées (4.1.2) des données socio-démographiques du Recensement de 2016 par Statistique Canada ne sont pas disponibles. Ces valeurs manquantes concernent une quinzaines de foyers qui n'ont pas été pris en compte dans les analyses. Par ailleurs, pour une AD, la valeur médiane des logements était de 0$. Par souci de cohérence elle a été ainsi remplacée par la valeur moyenne des logements de cette AD.

4.1.4 Création de sous-ensembles

Les inondations du printemps 2019 ont touché différentes parties du Québec, qui sont parfois très éloignées. Ainsi, la configuration de points des données de la CRC est localisée dans une vaste fenêtre d'observation, ce qui peut manquer de pertinence pour certaines

analyses. Par exemple, l'étude de l'intensité dans une fenêtre aussi large n'est pas vraiment adaptée, dans la mesure où il est relativement aisé d'observer que les points ne sont pas répartis uniformément sur la carte (voir figure 4.2). On peut alors créer des sous-ensembles de points à partir du jeu de données initial, pour étudier des configurations au sein de fenêtres d'observations plus petites. Les jeux de données *Montreal* et *Marthe* ont ainsi été créés en sélectionnant les foyers dans les zones respectives du Grand Montréal (approximativement) et de la ville de Sainte-Marthe-sur-le-Lac. La zone *Marthe* est située à l'intérieur de la zone *Montreal* et contient environ un quart des observations de jeu de données complet du Québec. La sélection des observations s'est effectuée selon le nom de la subdivision de recensement de chaque point comme le détaille le tableau 4.2.

Zone	Noms des subdivisions de recensement
Montreal	Sainte-Marthe-sur-le-Lac, Laval, Montréal, Oka, Hudson, L'Île-Perrot, Rigaud, Saint-Placide, Saint-André-d'Argenteuil, Saint-Colomban, Lachute, Pointe-Calumet, Terrasse-Vaudreuil, Mirabel, Vaudreuil-Dorion, Vaudreuil-sur-le-Lac, Rosemère, Pincourt, Deux-Montagnes, Saint-Joseph-du-Lac, Boisbriand, Saint-Eustache, Terrebonne, Senneville, Pointe-Fortune, Kanesatake, Léry, Saint-Constant
Marthe	Sainte-Marthe-sur-le-Lac

TABLEAU 4.2 – Noms des subdivisions de recensement utilisées pour sélectionner les foyers pour les sous-ensembles de points *Montreal* et *Marthe*.

Ainsi, dans les sections suivantes, la zone *Marthe* fera référence aux 1 222 observations situées dans la ville de Sainte-Marthe-sur-le-Lac, la zone *Montréal* à celle des 2 406 observations du Grand Montréal et la zone *Québec* concernera le jeu de données au complet (4 705 foyers).

4.1.5 Analyse descriptive des variables

Afin de mieux appréhender les futures analyses, il est important d'étudier de manière descriptive les différentes variables disponibles dans le cadre de cette étude. Cette section a donc pour but de présenter les éléments principaux pour chaque variable. L'annexe A complète ces informations avec d'autres détails liés à la distribution des variables. Les

tableaux des statistiques descriptives pour les zones de *Montréal* et *Marthe* se trouvent à l'annexe A.

La variable d'intérêt principal pour cette étude, le montant total alloué au foyer (ou *montant reçu*) en $ CAD, possède différentes caractéristiques nécessaires à la compréhension des futures analyses et de leurs résultats. Pour les 4 705 observations (ou foyers) sélectionnées après le pré-traitement des données, il y a 1 196 montants uniques qui ont été versés par la CRC. En d'autres termes, de nombreux foyers ont reçu des sommes identiques, ce qui indique une faible variabilité dans la variable d'intérêt Y (*montant reçu*). En effet, 2 145 foyers ont reçu une aide égale à 600$, ce qui correspond à plus de 45% des observations du jeu de données *Québec*. Ce pourcentage n'est pas surprenant car il s'aligne avec la politique d'aide financière apportée par la CRC. De manière générale, elle se compose de plusieurs programmes d'assistance, appelés Directives, qui ont des objectifs distincts. Dans le cadre du programme d'assistance financière à grande échelle (*Mass Assistance*) faisant partie de ceux ayant été déployés au Québec pour les inondations, l'objectif est d'aider le plus de sinistrés dans un délai très court. Cette aide est basée sur un principe de neutralité et des montants uniques sont alors alloués le plus rapidement possible aux foyers ayant été impactés afin de les aider à couvrir les dépenses imprévues d'évacuation.

Concernant la répartition des montants, elle s'étend de 0$ (37 foyers) à 33 506$. La moyenne du total des versements pour l'ensemble du jeu de données est de 1 437$, cependant la médiane est plus modeste et se situe à 600$ (voir tableau 4.3). Pour les sous-ensembles *Montréal* et *Marthe*, qui représentent respectivement 51% et 26% du jeu de données *Québec* (jeu de données complet), ces valeurs sont plus élevées (voir les tableaux 1 et 2 à l'annexe A).

Le diagramme à moustache (figure 4.3) présente la répartition du nombre de personnes dans les foyers selon différentes catégories. On constate que la répartition des hommes et des femmes dans les foyers est très similaire, avec des médianes égales à 1 ainsi que des moyennes de 1,10 et 1,11. Concernant les catégories d'âge, c'est sans surprise celle des adultes qui est la plus représentée, suivie par les personnes de plus de 65 ans. Le tableau 3

de l'annexe A présente les fréquences des tranches d'âge pour le jeu de données complet. Au total, un foyer comprend en moyenne 2,2 personnes et peut aller jusqu'à 9 personnes.

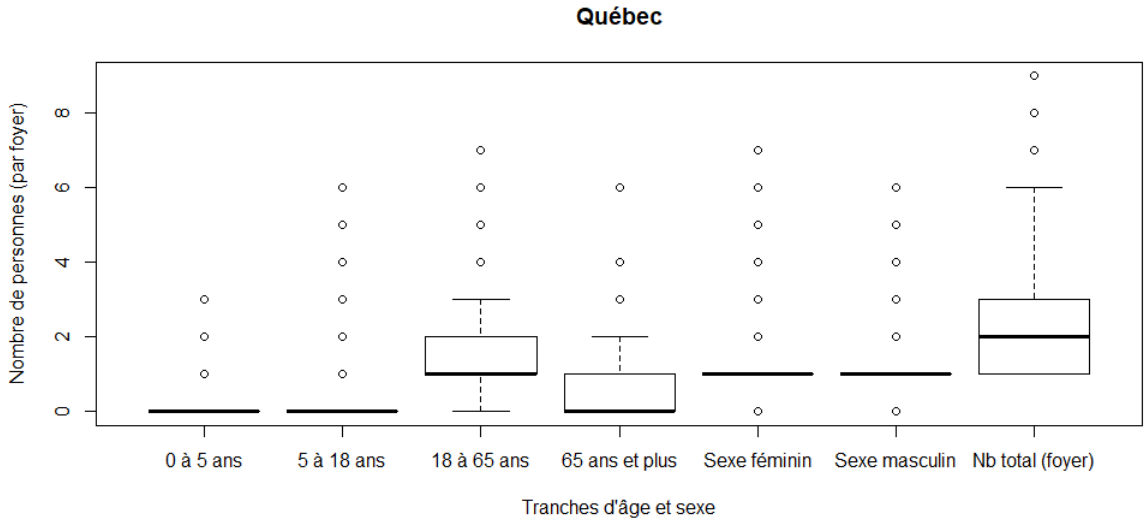


FIGURE 4.3 – Distribution selon les catégories d'âge et de sexe du nombre de personnes dans les foyers pour le jeu de donnée complet *Québec*.

Pour les variables externes issues du recensement de 2016 de Statistique Canada, le tableau 4.3 présente aussi leur différentes distributions. On constate de grandes disparités parmi certaines d'entre-elles avec par exemple le revenu total médian des ménages, où la valeur minimale est de 20 064$ et la valeur maximale est de 173 568$. Quant à la variable de sévérité des inondations, la catégorie « extrême » est la plus représentée. Ainsi, parmi les 4 705 foyers, 3 784 d'entre-eux sont situés au plus proche d'une zone ayant subi des inondations de la plus haute intensité. Près de 93% des foyers situés dans le sous-ensemble de données *Marthe* sont au plus proche d'une zone de sévérité « extrême ». La figure 4.4 représente la distribution du montant reçu en fonction du niveau de sévérité associé aux foyers sinistrés. Ainsi, on observe que l'étendue inter-quartile augmente à mesure où la sévérité s'aggrave et la médiane du montant reçu est plus élevée pour la catégorie « extrême » que pour les trois autres catégories.

En ce qui concerne la corrélation, qui mesure la relation linéaire entre deux variables, on observe une faible corrélation entre le montant reçu et les autres variables. La figure

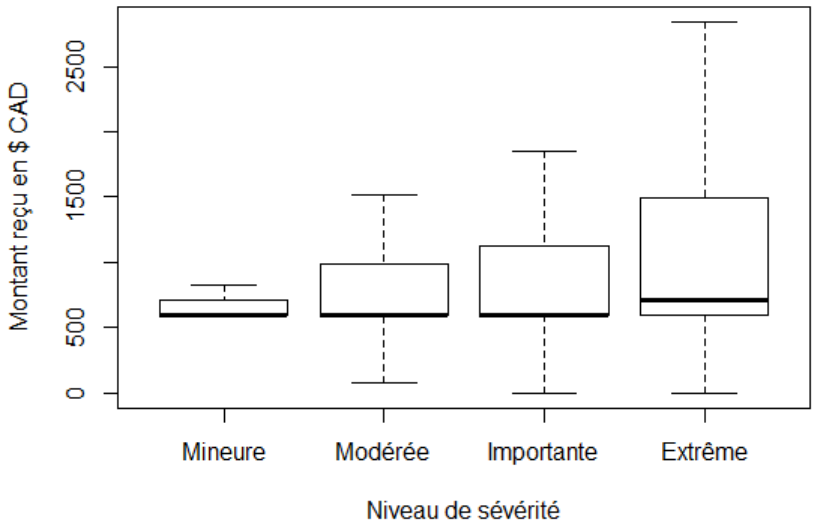
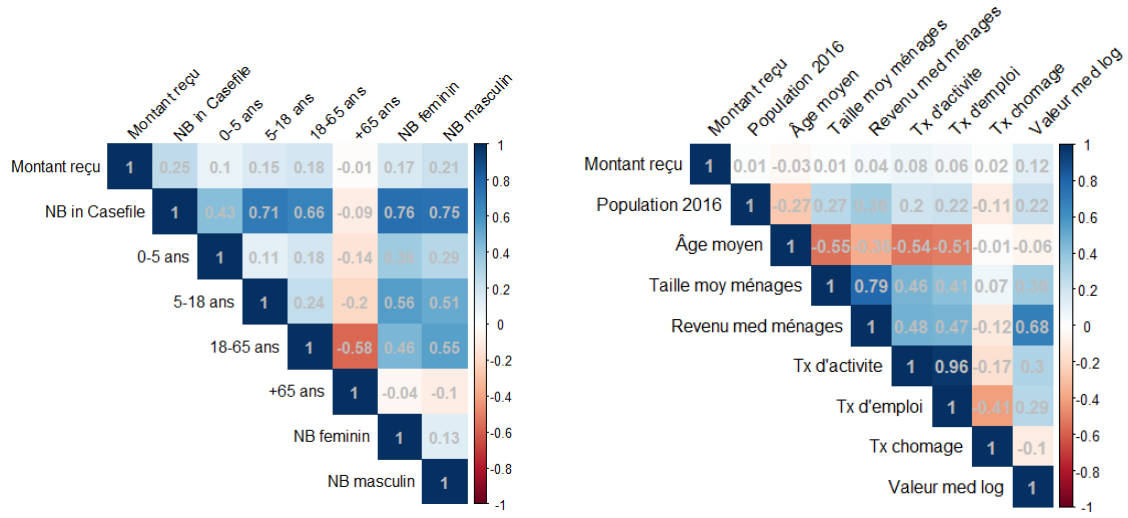


FIGURE 4.4 – Distribution du montant d'argent reçu par foyer selon le niveau de sévérité. Les valeurs extrêmes n'ont pas été prises en compte pour ce diagramme.

4.5 présente ces niveaux de corrélation où l'on remarque que la variable la plus corrélée avec le montant reçu est celle du nombre de personnes dans le foyer avec une corrélation positive de 0,25.

Pour conclure, à elles seules, les variables présentées ne semblent pas posséder un grand potentiel pour expliquer la variabilité du montant alloué aux familles. La pertinence de l'utilisation des coordonnées spatiales dans les modèles d'analyses est alors une excellente possibilité à explorer. Les sections suivantes auront ainsi pour but d'exploiter les informations géographiques fournies par la CRC en appliquant les méthodes présentées aux chapitres 2 et 3 afin d'enrichir les analyses et leurs résultats. Ces analyses nous permettront de détecter si l'utilisation de l'information spatiale est pertinente dans le cadre de la valorisation des données recueillies par la Croix-Rouge canadienne à la suite des inondations de la province du Québec en 2019.



(a) Données CRC (b) Données Statistique Canada

FIGURE 4.5 – Corrélations entre le montant reçu par foyer (*Montant reçu*) et les autres variables disponibles.

4.2 Analyse des données ponctuelles

Cette section a pour objectif d'appliquer les méthodes d'analyse des données ponctuelles (chapitre 2) sur les données relatives à cette étude de cas.

4.2.1 Estimation de l'intensité par la méthode des quadrats

Comme présenté au chapitre 2, l'utilisation des quadrats est une manière simple de vérifier l'hétérogénéité d'un processus et de regarder si les régions de même superficie contiennent environ le même nombre de points (comme cela devrait être le cas pour un processus d'intensité homogène). Dans le cas des données de la CRC, il semble évident que le jeu de données complet du Québec ne présente pas une intensité homogène. Il semble plus pertinent d'étudier la nature de l'intensité dans les zones du Grand Montréal ou de Sainte-Marthe-sur-le-Lac. Sans grande surprise, on peut constater que pour le jeu de données de la zone du Grand Montréal, le quadrat qui se distingue avec une intensité beaucoup plus forte que les autres est celui de la zone de Sainte-Marthe-sur-le-Lac (voir la figure 4.6). Lorsque l'on se concentre uniquement sur la zone du jeu de données *Marthe*,

Variable	Min	1er Q	Med	Moy	3ème Q	Max
Données de la Croix-Rouge canadienne						
Montant reçu ($ CAD)	0	600	600	1 399	1 300	33 506
NB de personnes	1	1	2	2,2	3	9
NB 0 à 5 ans	0	0	0	0,11	0	3
NB 5 à 18 ans	0	0	0	0,33	0	6
NB 18 à 65 ans	0	1	1	1,40	2	7
NB + de 65 ans	0	0	0	0,38	1	6
NB sexe féminin	0	1	1	1,11	1	7
NB sexe masculin	0	1	1	1,10	1	6
Données externes						
Population 2016	259	506	596	759,2	785	5 210
Âge moyen	31,2	40,5	43.6	43,4	45,9	69,2
Taille moy ménages	1,5	2,1	2,3	2,3	2,4	3,2
Revenu med ménages	20 064	41 499	61 312	61 093	69 248	173 568
Taux d'activité	30,2	58,1	62,6	63,7	69,9	85,3
Taux d'emploi	26,7	54,8	57,1	59,4	65,2	78,5
Taux de chômage	0	4,5	6,7	7,0	8,3	21,4
Valeur med logements	100 192	180 219	200 669	218 194	250 554	898 905
	Fréquences (sur 4705 foyers)					
Sévérité	Mineure	Modérée	Importante		Extrême	
	57	271	593		3 784	
	(1.2%)	(5.8%)	(12.6%)		(80.4%)	

TABLEAU 4.3 – Statistiques descriptives des variables disponibles pour le jeu de données complet (Québec) de 4705 observations après pré-traitement.

l'analyse des quadrats est plus pertinente car ils présentent des intensités différentes sans devoir choisir des dimensions très petites. De plus, l'hypothèse d'une intensité homogène dans les trois zones d'études est bien écartée par le test d'homogénéité selon la méthode des quadrats. Les tests effectués rejettent tous l'hypothèse nulle (processus de Poisson homogène) avec un niveau de significativité de 1%.

Estimation de la densité par la méthode du noyau

Généralement, le choix du noyau impacte peu les résultats du lissage. Ici, le choix du noyau n'a pas un impact très important sur l'estimation de la densité. La figure 4.8 montre qu'en effet, les différences sont minimes et peu perceptibles entre les différents noyaux

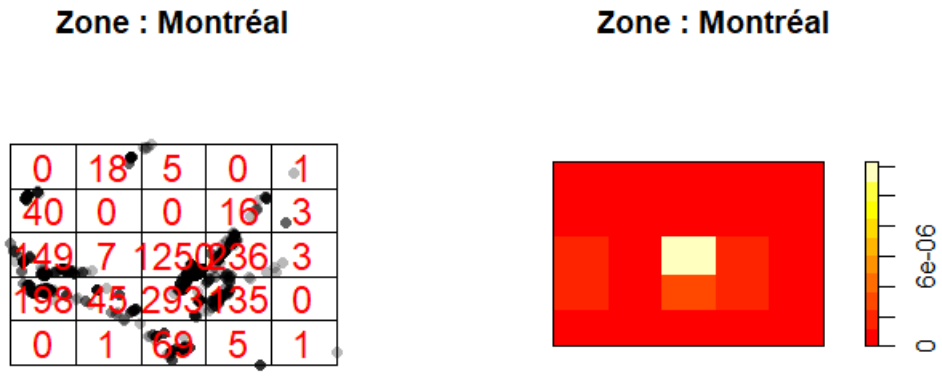


FIGURE 4.6 – Analyse de l'intensité par la méthode des quadrats (zone de Montréal). À gauche : dénombrement des foyers situés dans chaque quadrat. À droite : mesure d'intensité (points par mètre carré).

proposés par *spatstat* pour la région de Sainte-Marthe-sur-le-Lac. Nous arrivons à cette même conclusion pour l'analyse du jeu de données du Québec et de Montréal. De manière similaire, le choix de la correction des effets de bords n'affecte pas vraiment les résultats. On privilégie cependant la correction de Diggle, car elle semble plus appropriée dans la mesure où la fenêtre d'observation n'est pas indépendante du processus sous-jacent.

Comme présenté par la figure 4.8, qui compare le lissage selon le type de noyau utilisé, on retrouve bien des niveaux d'intensité semblables à ceux présentés par la méthode des quadrats (figure 4.7).

Concernant le choix de la bande passante, les différences sont en effet plus marquées. Une bande passante avec un rayon élevé conduira à une densité fortement lissée, et donc un biais élevé. À l'inverse, la densité sera moins lissée avec un plus petit rayon. Il est donc pertinent de tester différentes bandes passantes ainsi que différentes méthodes pour les optimiser. Différentes méthodes permettent de sélectionner automatiquement une bande passante selon l'optimisation de certains critères. Par exemple, une des méthodes pro-

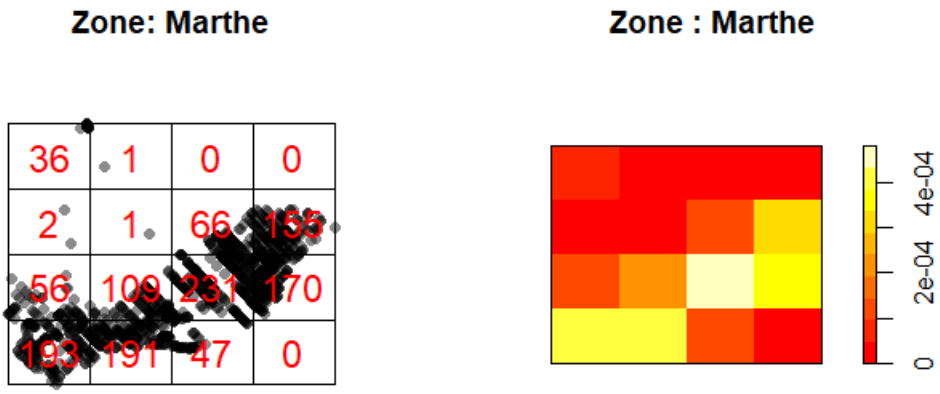


FIGURE 4.7 – Analyse de l'intensité par la méthode des quadrats (zone de Sainte-Marthe-sur-le-Lac). À gauche : dénombrement des foyers situés dans chaque quadrat. À droite : mesure d'intensité (points par mètre carré).

Méthodes	σ (en mètres)		
	Quebec	Montreal	Marthe
`bw.diggle`	109.18	10.19	10.19
`bw.ppl`	1751.30	475.21	34.34
`bw.frac`	93986.69	16602.62	908.87
`bw.scott`	30641.83 (x) 10431.30(y)	3078.39(x) 1265.80(y)	242.73(x) 142.68(y)

TABLEAU 4.4 – Bandes passantes optimales ($h = 2\sigma$) selon différentes méthodes disponibles dans *spatstat*.

posées par Diggle a pour but de choisir la bande passante qui minimise l'erreur quadratique moyenne de l'estimateur. Une autre méthode se base sur le calcul d'un estimateur de maximum de vraisemblance en utilisant la validation croisée pour choisir la bande passante optimale, soit la fonction `bw.ppl()` de *spatstat*. Le tableau 4.2.1 présente les différentes bandes passantes optimales sélectionnées selon quatre méthodes disponibles dans le package *spatstat*.

On remarque ainsi que les méthodes 3 et 4 ont des bandes passantes extrêmement plus

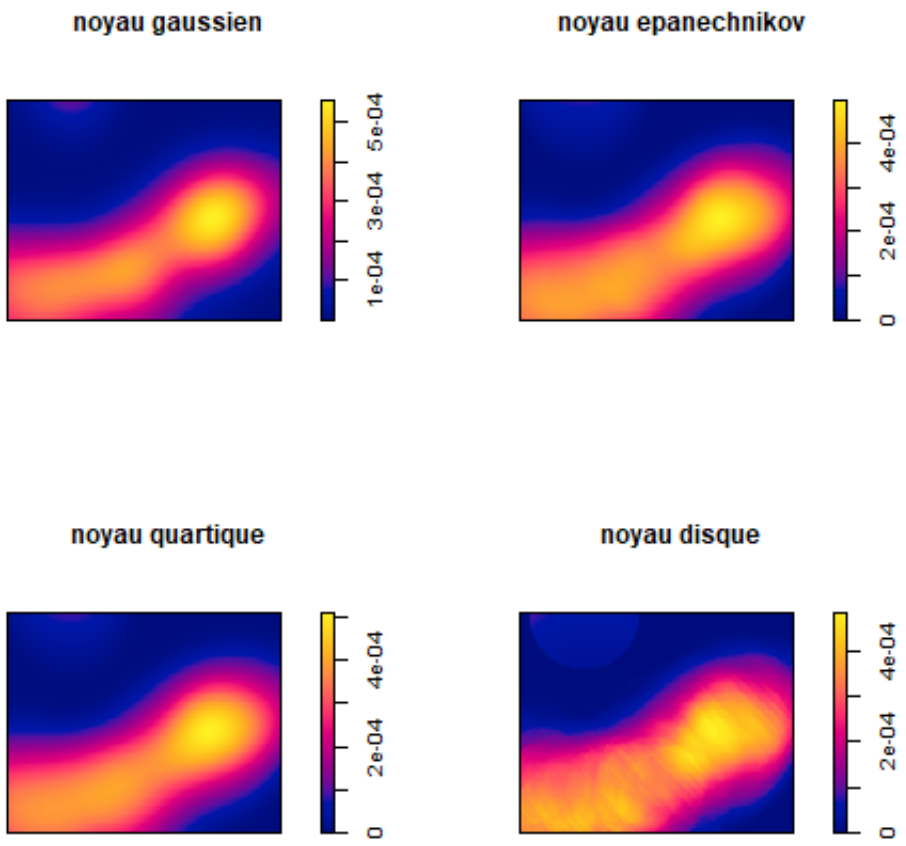


FIGURE 4.8 – Estimation de la densité de la zone *Marthe* selon plusieurs noyaux.

élevées que celles des méthodes 1 et 2. Pour la méthode 4, qui est basée sur la règle de Scott (1992), les résultats de la bande passante se lisent sous la forme d'un vecteur et non d'un unique rayon : 30 641,83 mètres dans la direction des x et 10 431,30 dans la direction des y pour le jeu de données *Quebec*. Il est à noter que les valeurs présentent dans le tableau 4.2.1 sont celles retournées par les fonctions présentent dans *spatstat* et qui correspondent au paramètre *sigma* (rayon de lissage $= 2 * sigma$). Pour avoir le réel rayon de la bande passante, il faut multiplier les résultats obtenus par deux.

4.2.2 Étude de la dépendance avec la fonction *K* de Ripley

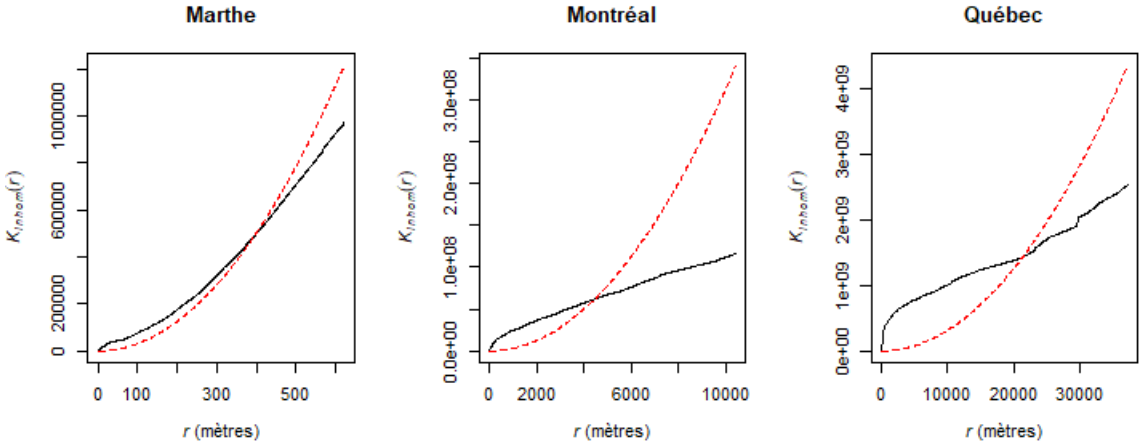


FIGURE 4.9 – Représentation des Fonctions K_{inhom} (courbes noires) pour les trois zones étudiées (Sainte-Marthe-sur-le-Lac, Grand Montréal et la province du Québec). Les courbes rouges en pointillés sont les valeurs de référence πr^2 pour tous les rayons r.

La figure 4.9 présente les fonctions $\hat{K}_{\text{inhom}}$ qui prennent en compte les intensités non-homogènes des trois configurations étudiées. Dans la plus petite fenêtre d'étude (*Marthe*), la courbe de la fonction $\hat{K}_{\text{inhom}}$ est assez proche de celle de référence, alors que pour les autres zones l'écart est plus important. À partir des rayons de 6 km pour Montréal et de 20 km pour le Québec, la dépendance entre les points décrite par les courbes s'inverse : on passe d'un processus agrégé à un processus régulier. Ces résultats semblent cohérents avec le contexte car, comme nous avons pu le voir par exemple sur la carte de la figure 4.2, la fenêtre d'observation pour la province du Québec est très grande et on y observe certaines zones d'agglomération de points mais aussi beaucoup de zones sans aucun point. Il est donc normal de s'attendre à ce qu'à partir d'une certaine distance, les observations se raréfient ce qui se traduit par une $\hat{K}_{\text{inhom}}$ en dessous de la courbe de référence mais qui ne signifie pas en réalité une répulsion entre les points dans ce contexte.

Permettant aussi de prendre en compte la non-homogénéité de l'espace, la fonction *D* (Diggle et Chetwynd, 1991) aurait pu être un outil intéressant à utiliser si les données spatiales relatives à l'ensemble des foyers sinistrés étaient disponibles (pas uniquement

ceux ayant reçu de l'aide de la part de la CRC). Ainsi, on aurait éventuellement pu étudier et détecter des écarts entre la distribution des foyers nécessitant l'aide de la CRC et la distribution complète des foyers touchés par les inondations (distribution de référence).

4.2.3 Analyse des marques

Dans cette section, nous effectuerons une analyse des marques selon les méthodes présentées à la section 2.3.

Statistiques descriptives et tendances spatiales

Comme mentionné à la section 2.3.1, une des premières étapes à effectuer en présence d'un processus marqué est une analyse descriptive pour chacune des marques présentes. Parmi les éléments principaux de cette analyse (effectuée à section 4.1.5) on trouve l'importante proportion (> 45%) de foyers ayant reçu exactement 600$ CAD, ce qui signifie une absence de variabilité dans la variable Y pour une grande partie des données. On constate aussi qu'un important pourcentage des observations (80,4%) est situé au plus proche d'une zone ayant été qualifiée d'une extrême sévérité selon *Données Québec*. Ici, les jeux de données (*Québec, Montréal et Marthe*) contiennent plusieurs marques associées aux points, dans la mesure où différentes variables ont été récoltées pour chaque foyer. Nous allons cependant nous intéresser uniquement à l'analyse du montant reçu, étant donné qu'il représente la marque principale de l'étude.

La carte de la figure 4.10 illustre la répartition géographique des montants pour l'intégralité du jeu de données *Québec*. Cependant à ce niveau, il est difficile d'observer une tendance spatiale bien définie. La carte de la figure 4.11 représente la zone de *Montréal*. Si à cette échelle il est plus aisé de distinguer la répartition individuelle des montants sur la carte, il reste néanmoins difficile de distinguer des tendances spatiales. L'utilisation de la méthode de lissage doit pouvoir faciliter cette identification.

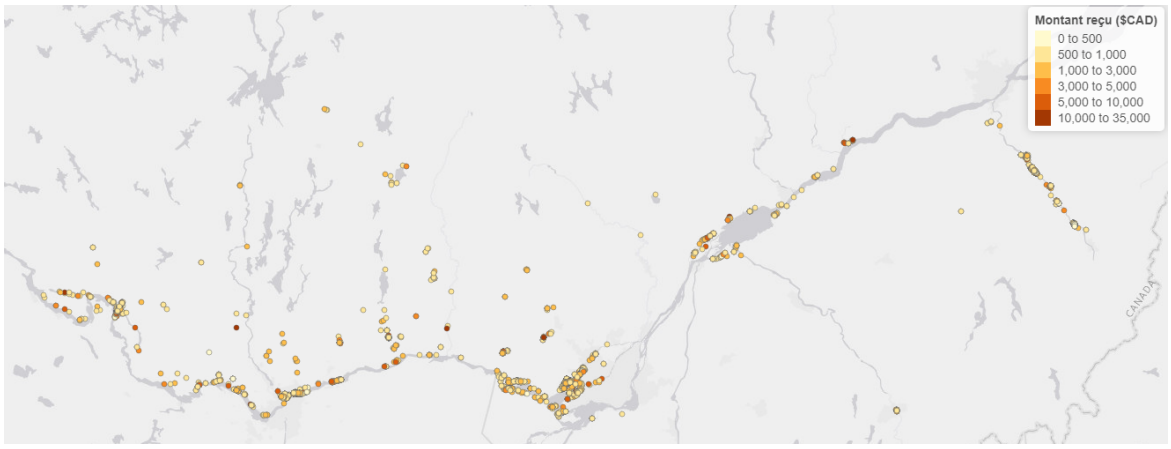


FIGURE 4.10 – Répartition spatiale des montants reçus en $ CAD pour le jeu de données *Québec*.

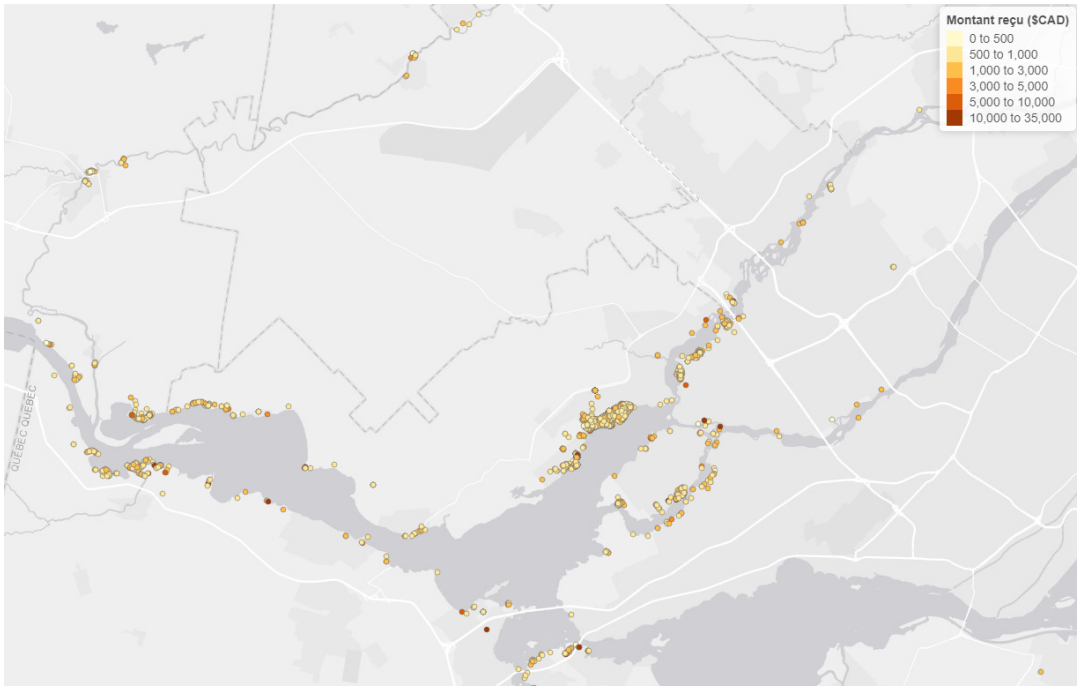
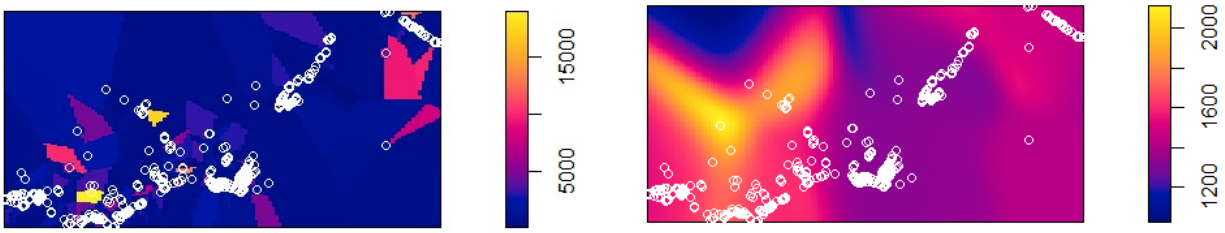


FIGURE 4.11 – Répartition spatiale des montants reçus en $ CAD pour le jeu de données *Montréal*.

Intensité des marques

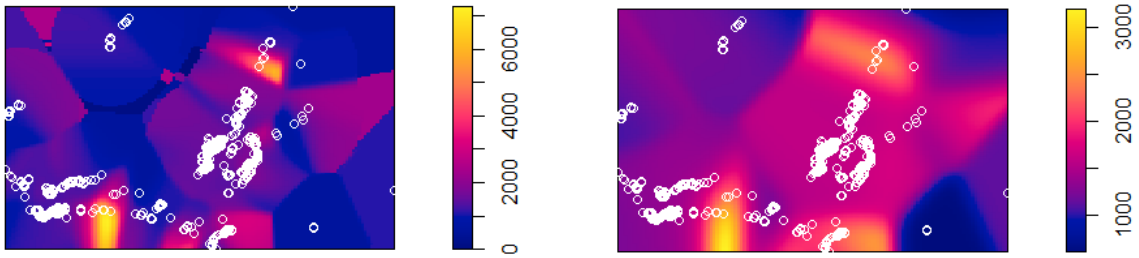
Concernant l'intensité des marques, les figures 4.12, 4.13 et 4.14 représentent la moyenne spatialement lissée $\tilde{m}(u)$ du montant reçu (en $ CAD) pour les différentes zones d'études.

La question qui nous intéresse ici est de savoir s'il y a une homogénéité spatiale de la valeur moyenne des montants reçus. En d'autres termes, est-ce que certains endroits ont reçu en moyenne plus de dons de la part de la CRC que d'autres localisations ? Même en faisant varier différents paramètres comme la taille de la bande passante ou la forme du noyau de la fonction `Smooth()` de *spatstat*, les résultats ne sont pas particulièrement concluants. De plus, on pourrait s'attendre à de meilleures conclusions avec une petite fenêtre d'observation comme celle de *Marthe*, mais elles restent difficiles à exprimer et tendent plutôt vers une moyenne uniforme dans cette zone.



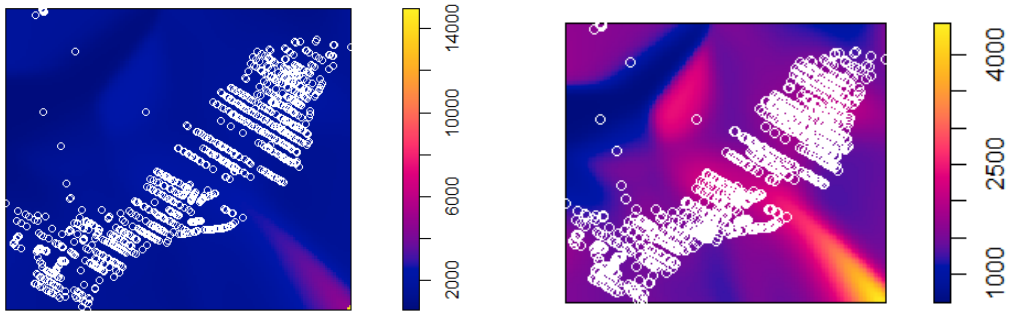
(a) Bande passante trop faible (b) Bande passante plus élevée

FIGURE 4.12 – Résultat du lissage de l'intensité de la marque du montant reçu en $ CAD pour *Québec* selon la taille des bandes passantes. Les cercles blancs représentent la localisation des observations.



(a) Bande passante trop faible (b) Bande passante plus élevée

FIGURE 4.13 – Résultat du lissage de l'intensité de la marque du montant reçu en $ CAD pour *Montréal* selon la taille des bandes passantes. Les cercles bancs représentent la localisation des observations.



(a) Bande passante trop faible (b) Bande passante légèrement plus élevée

FIGURE 4.14 – Résultat du lissage de l'intensité de la marque du montant reçu en $ CAD pour *Marthe* selon la taille des bandes passantes. Les cercles bancs représentent la localisation des observations.

Dépendance et corrélation

Concernant la dépendance et la corrélation entre les valeurs des marques, la fonction de corrélation de marque et la fonction K pondérée par la marque peuvent nous aider pour formuler des conclusions. Les graphiques de droite des figures 4.15, 4.16 et 4.17 représentent l'estimation de la fonction de corrélation (avec une correction isotropique) pour les différentes zones d'étude. La fonction de corrélation estimée $\hat{k}_{mm}$ est représentée par la courbe noire et la courbe rouge (pointillés) représente la valeur de référence suggérant une absence de corrélation. Pour l'ensemble du jeu de données (*Québec*) l'estimation de la fonction de corrélation suggère une corrélation plutôt faiblement négative, voire nulle, jusqu'à un rayon de 40 km (40 000 mètres). On observe un pic entre 40 km et 50 km de rayon, qui signifierait donc une plus forte corrélation entre les observations éloignées. Concernant les plus petites fenêtres d'observation de *Montréal* et *Marthe*, les conclusions sont aussi nuancées. Pour *Montréal*, on observe des variations entre corrélation négative et corrélation positive, mais la tendance générale semble être similaire à celle du Québec. En revanche, pour la zone de *Marthe*, la fonction d'estimation de corrélation semble indiquer une corrélation majoritairement positive entre les observations, à l'exception d'une association négative pour des rayons inférieurs à 50 m ou supérieurs à 450 m. Pour l'estimation de la fonction K pondérée, les résultats pour les trois jeux de données sont plus

clairs et tendent largement à indiquer la présence d'une agrégation, dans la mesure où les courbes d'estimations sont bien au dessus des courbes de références (suivant un processus de Poisson aléatoire).

Pour terminer, comme mentionné à la section 2.3.1, une alternative à ces fonctions est de considérer uniquement les voisins les plus proches avec le calcul de la corrélation entre la marque d'un point et la marque de son voisin le plus proche. Pour les trois zones, cette corrélation est très proche de 0 (*Québec* : $0,006$, *Montréal* : $-0,003$, *Marthe* : $-0,01$), ce qui ne permet pas de conclure à une corrélation entre les marques des plus proches voisins.

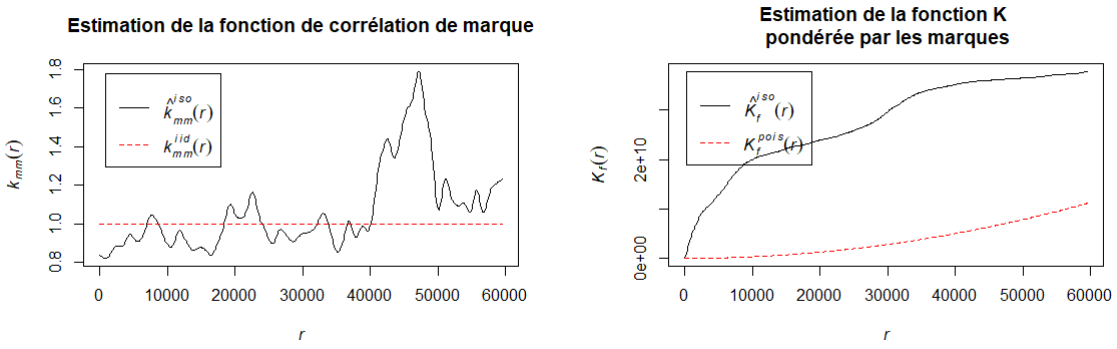


FIGURE 4.15 – Fonction de corrélation de marque (montant reçu) et fonction K pondérée par la marque (montant reçu) pour le jeu de données du *Québec*. La courbe noire représente la fonction estimée et la rouge est une valeur de référence.

4.2.4 Régression géographiquement pondérée

Pour l'ensemble des données du Québec, le tableau 4.5 présente les résultats de deux modèles de régression linéaire classique, où le premier modèle ne comporte que les variables récoltées par la CRC ainsi que la variable de sévérité (*Données Québec*) et où le second modèle comprend en plus les variables de Statistique Canada. Afin de ne pas avoir de relation linéaire parfaite entre certaines variables, *NB sexe masculin* et *NB 18 à 65 ans* ne font pas partie des variables présentes dans les modèles. On constate que l'ajout des variables de Statistique Canada permet d'obtenir un meilleur R^2 ajusté, qui reste cepen-

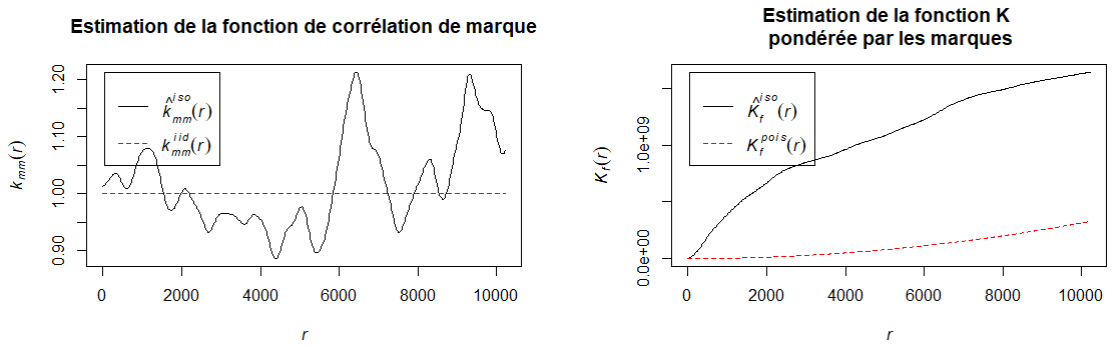


FIGURE 4.16 – Fonction de corrélation de marque (montant reçu) et fonction K pondérée par la marque (montant reçu) pour le jeu de données du *Montréal*. La courbe noire représente la fonction estimée et la rouge est une valeur de référence.

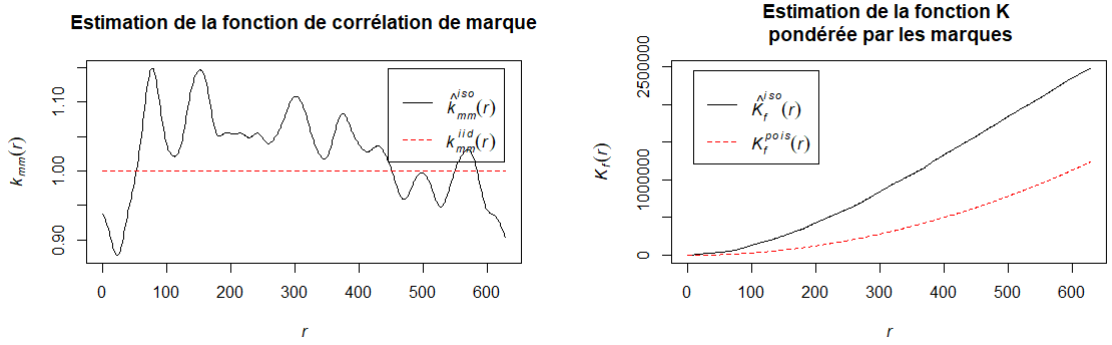


FIGURE 4.17 – Fonction de corrélation de marque (montant reçu) et fonction K pondérée par la marque (montant reçu) pour le jeu de données du *Marthe*. La courbe noire représente la fonction estimée et la rouge est une valeur de référence.

dant très faible. Ainsi, la proportion de la variabilité du montant reçu (Y) expliqué par les variables de la CRC et de la sévérité est de 6,9%. Cette proportion augmente à 9,7% lorsque que le modèle inclut les variables de Statistique Canada.

Comme présenté à la section 2.4, la régression géographiquement pondérée permet de prendre en compte les variations locales au sein des données, contrairement à la régression linéaire classique qui considère les coefficients comme uniques dans toute la zone d'étude. Il peut ainsi être intéressant d'étudier les résultats d'une régression géographiquement pondérée, qui inclurait alors les informations de localisation récoltées par la CRC. Dans un premier temps, le tableau 4.6 présente les performances des RGP, se-

Modèles	Coefficients	Estimations	Valeur-p	Performance
1	(Constante)	205,57	0,440	R^2 adj = 0,069 ~ Valeur-p du test Global = $< 0,001$
	NB in Casefile	522,51	$< 0,001$	
	NB sexe féminin	-119,13	0,018	
	NB 0 à 5 ans	-129,19	0,137	
	NB 5 à 18 ans	-126,93	0,025	
	NB + de 65 ans	-5,62	0,901	
	Sévérité Modérée	-69,34	0,808	
	Sévérité Importante	28,67	0,922	
	Sévérité Extrême	285,03	0,276	
2	(Constante)	895,4	0,260	R^2 adj = 0,097 ~ Valeur-p du test Global = $< 0,001$
	NB in Casefile	508,3	$< 0,001$	
	Sévérité Modérée	296,7	0,302	
	Sévérité Importante	223,5	0,412	
	Sévérité Extrême	571,3	0,031	
	NB sexe féminin	-100,1	0,043	
	NB 0 à 5 ans	-93,9	0,273	
	NB 5 à 18 ans	-109,4	0,050	
	NB + de 65 ans	-1,4	0,974	
	Population2016	0,02	0,717	
	Âge moyen	-22,4	0,010	
	Taille moy ménages	-393,3	0,045	
	Revenu med ménages	-0,02	0,001	
	Tx d'activité	135,3	0,001	
	Tx d'emploi	-129,2	0,001	
	Tx de chômage	-56,1	0,021	
	Valeur med logements	0,006	$< 0,001$	

TABLEAU 4.5 – Résultats des modèles de régression linéaire classique selon les variables utilisées. Note : pour Séverité, la valeur de référence est Sévérité *Faible*.
Les variables *NB sexe masculin* et *NB 18 à 65 ans* ne sont pas présentes dans le modèle et servent ainsi de valeur de référence.

lon les noyaux, des deux modèles utilisés précédemment dans le cadre de la régression linéaire classique. Lorsque l'on compare la performance des différents modèles de RGP, on constate que l'utilisation d'une fonction noyau exponentiel fixe permet d'obtenir une valeur de R^2 ajusté de 19,2% et un AIC de 84 348 (en incluant toutes les variables). Ces mesures montrent que ce modèle semble mieux performer que les autres pour le jeu de données complet du Québec. En choisissant ces paramètres d'estimation, la prise en compte des variations locales par la RGP permet d'améliorer le R^2 ajusté, qui augmente de 6,9% à 17,6% pour le modèle n'incluant pas les données socio-démographiques (StatCan) et de 9,7% à 19,2% pour le modèle les incluant. Si les résultats des modèles de RGP sont meilleurs que ceux de la régression linéaire, la variabilité du montant alloué reste mal expliquée par les données (19,2%), même en incluant l'information géographique disponible. L'utilisation des variables de Statistique Canada améliore légèrement les performance du modèle (d'un R^2 de 17,6% à 19,2%) dans le cas du noyau exponentiel fixe.

Il est aussi intéressant de remarquer que lorsque les noyaux sont adaptatifs, les bandes passantes des noyaux Bicarré et Tricube sont élevées et proches du nombre total d'observations (4 705). Cela peut indiquer que l'hétérogénéité locale n'est potentiellement pas significative et qu'une RGP n'est probablement pas très pertinente.

Les statistiques descriptives des coefficients de ces deux modèles de RGP sont disponibles dans le tableau 4.7. On distingue aisément qu'il existe de grandes différences dans les valeurs des coefficients pour une même variable. Dans le cas de la variable *NB in Casefile*, qui correspond au nombre de personnes dans le foyer, le coefficient s'étend de $-668,5$ à 4 334,5. Cela signifie que pour certaines localisations, selon ce modèle, le nombre de personnes dans le foyer peut avoir un impact négatif sur le montant versé par la CRC. La figure 4.18 permet de visualiser la distribution et la répartition géographique du coefficient de régression de la variable *NB in Casefile* pour le modèle de RGP numéro 2 (incluant les variables de StatCan). Comme on peut le constater sur la boîte à moustaches (figure 4.18(b)), il y a peu de valeurs négatives pour ce coefficient. La figure 4.18(a) permet de détecter les lieux où les valeurs du coefficient sont négatives (ou faibles) pour la variable *Nb in Casefile* et, à l'inverse, où elles sont les plus élevées. De manière simi-

laire la figure 4.19 représente l'estimation continue du coefficient pour la variable *NB in Casefile*.

Noyau		*Modèle 1 : variables CRC + sévérité*		*Modèle 2 : CRC + sévérité + StatCan*	
		Bande passante	*Performance*	*Bande passante*	*Performance*
Gaussien	Fixe	13 903,6	R^2 adj = 0,124 AIC= 84 511	68 865,5	R^2 adj = 0,107 AIC = 84 531
	Adaptatif	2 437	R^2 adj = 0,072 AIC = 84 678	2 452	R^2 adj = 0,102 AIC = 84 538
Exponentiel	Fixe	5 902,4	R^2 adj = 0,176 AIC = 84 331	7 560,7	R^2 adj = 0,192 AIC = 84 348
	Adaptatif	94	R^2 adj = 0,134 AIC = 84 547	2 476	R^2 adj = 0,103 AIC= 84 490
Bicarré	Fixe	164 125.4	R^2 adj = 0,075 AIC = 84 674	177 098,2	R^2 adj = 0,102 AIC = 84 534
	Adaptatif	3 145	R^2 adj = 0,086 AIC = 84 697	3 509	R^2 adj = 0,108 AIC = 84 520
Tri-cube	Fixe	169 026	R^2 adj = 0,075 AIC = 84 676	182 087	R^2 adj = 0,106 AIC = 84 535
	Adaptatif	3 145	R^2 adj = 0,077 AIC =84 665	3 509	R^2 adj = 0,108 AIC = 84 523
Boxcar	Fixe	298 109	R^2 adj = 0,069 AIC = 84 693	196 415	R^2 adj = 0,102 AIC = 84 547
	Adaptatif	1 375	R^2 adj = 0,076 AIC = 84 677	1 394	R^2 adj = 0,107 AIC = 84 543

TABLEAU 4.6 – Performances des modèles selon les valeurs des bandes passantes optimisées pour le jeu de données du Québec avec la fonction `bw.gwr()` de *GWmodel*. Lorsque l'option `adaptive = TRUE`, la valeur de la bande passante est exprimée en nombre de points voisins et non en termes de distance.

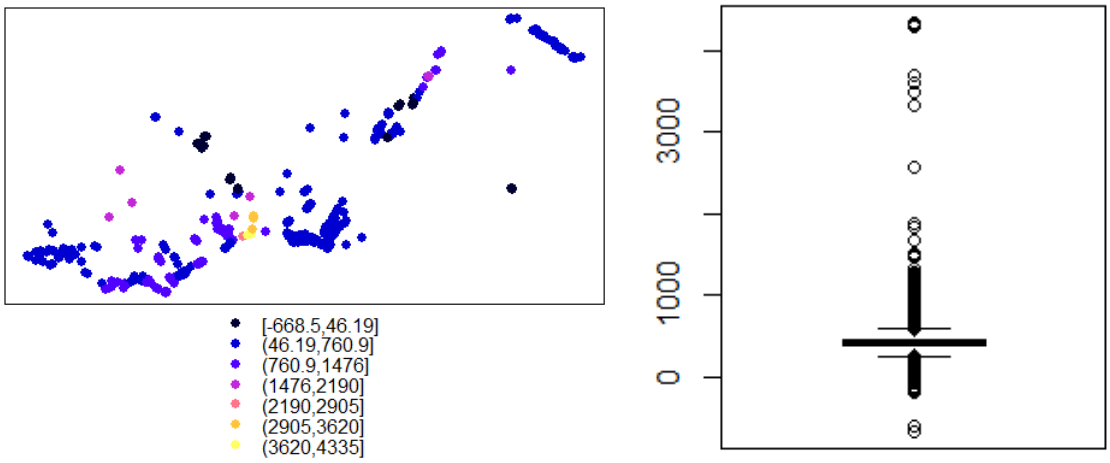
Pour les jeux de données représentant les zones *Montréal* et *Marthe*, qui comprennent moins d'observations, les performances des modèles sont plus faibles que pour la zone de *Québec*. De manière similaire à ce qui a été proposé pour la zone de *Québec*, l'annexe 4 présente les résultats obtenus pour les zones de *Montréal* et *Marthe*. Si l'ajout des variables de Statistique Canada a permis d'augmenter légèrement les performances du modèle pour les données du *Québec*, c'est aussi le cas pour *Montréal* ou *Marthe*. Cependant,

Modèle	Coefficients	Min	1er Q.	Med	3ème Q	Max
	(constante)	−5 509,7	−125,0	457,4	566,4	1 971,3
	NB in Casefile	−923,3	396,6	447,6	520,3	7 783,5
	NB sexe femme	−7 969,3	−85,8	1,4	52,4	2 217,3
Modèle 1 :	*NB 0 à 5 ans*	−7 675,3	−343,6	−174,8	−114,9	15 810,0
R^2 adj = 0,176	*NB 5 à 18 ans*	−7 371,6	−233,6	−183,6	−55,1	1 453,2
AIC = 84 331	*NB + de 65 ans*	−2 992,4	−64,6	−32,0	3,6	995,4
	Sévérité Modérée	− 2 969,4	−448,8	−274,8	246,6	16 450,3
	Sévérité Importante	−2 214,6	−374,2	−234,1	575,1	1 887,6
	Sévérité Extrême	−1 502,1	−150,4	−26,2	722,6	2 638,4
	(constante)	−16 743	−260,5	410,5	2 610,1	26 247
	NB in Casefile	−668,5	375,7	402,7	460,1	4 334,5
	NB sexe femme	−1 693,4	−75,3	22,4	79.2	1 657,0
	NB 0 à 5 ans	−5 646,0	−313,7	−114,9	−65,1	15 586,0
	NB 5 à 18 ans	−5 150,0	−227,3	−132,4	−59,7	1 267,0
	NB + de 65 ans	−2 438,5	−47,4	−21,7	12,8	902,4
	Sévérité Modérée	−2 817,7	−60,5	258,3	419,0	9 108,1
Modèle 2 :	*Sévérité Importante*	−2 094,6	−63,9	143,9	487,0	2 571,9
R^2 adj = 0,192	*Sévérité Extrême*	−1 097,5	292,3	530,8	631,6	3 186,2
AIC = 84 348	*Population 2016*	−1,3	−0,04	−0,03	0,04	1,8
	Âge moyen	−329,4	−18,3	−16,2	1,9	150,4
	Taille moy ménages	−3 647,2	−767,8	−355,3	−299,0	6 040,2
	Revenu med ménages	−0,20	−0,02	−0,009	−0,007	0,03
	Taux d'activité	−818,5	87,4	99,3	127,4	1 929,5
	Taux d'emploi	−202,7	−116,7	−93,6	−58,6	909,0
	Taux chômage	−1 394,6	−45,1	−37,9	−17,8	536,5
	Valeur med log	−0,009	0,003	0,005	0,006	0,016

TABLEAU 4.7 – Statistiques descriptives des coefficients des RGP (noyau exponentiel fixe) avec les performances des modèles associés.

ces zones ayant moins de diversité dans les valeurs de ces variables (voir pré-traitement à la section 4.1.3), il devient moins pertinent d'appliquer une RGP avec ces variables. En effet, les différents noyaux ne permettent pas toujours d'obtenir de meilleures résultats en comparaison d'une régression linéaire classique.

En conclusion, l'utilisation de modèles de régression géographiquement pondérée nous permet d'obtenir de meilleures mesures de performance pour le jeu de données complet (*Québec*) en comparaison aux modèles de régression linéaire classiques, cependant



(a) Répartition géographique des valeurs (b) Distribution des valeurs du coefficient

FIGURE 4.18 – Répartition géographique et distribution des valeurs du coefficient de RGP de la variable *NB in Casefile* (nombre de personne dans le foyer sinistré). Modèle 2 avec noyau exponentiel fixe (voir tableau 4.7).

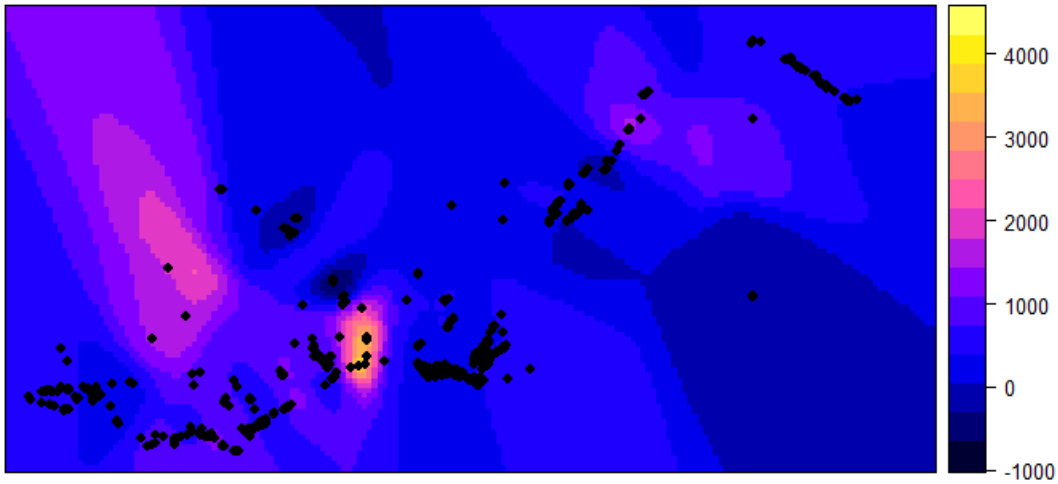


FIGURE 4.19 – Estimation continue du coefficient *NB in Casefile* pour le modèle de RGP avec noyau exponentiel fixe, sur un carroyage avec une précision de 9 km^2.

cette amélioration reste mineure. En effet, dans le cadre de la RGP, les variables disponibles ne permettent d'expliquer qu'une petite partie de la variation du montant alloué par la Croix-Rouge canadienne à chaque foyer sinistré. Par ailleurs, on a aussi observé que ces résultats ne sont pas forcément robustes à des légères modifications dans le jeu de

donnée étudié. Il faut donc faire preuve de prudence quant à leurs interprétations.

4.3 Analyse des données surfaciques

Dans cette section, nous allons appliquer les méthodes d'analyses des données surfaciques (voir chapitre 3) aux données récoltées par la CRC ainsi qu'aux données externes. Cette section a donc pour but d'étudier la structure de voisinage des observations puis de quantifier et modéliser l'influence qu'exercent les régions sur leurs voisines pour l'allocation des dons de la Croix-Rouge canadienne.

4.3.1 Description des données

Pour commencer, le fichier des limites des aires de diffusion mis à disposition par Statistique Canada est un élément central de cette analyse. Comme détaillé à la section 4.1.3, chaque foyer sinistré est associé à l'aire de diffusion lui correspondant. Il est ensuite possible d'agréger les données récoltées par la CRC au niveau des AD. La figure 4.20 représente la visualisation de 290 aires de diffusion associées aux foyers sinistrés enregistrés par la CRC. Cependant, parmi ces 290 AD, de nombreux polygones ne sont liés qu'à très peu d'observations. À titre d'exemple, 142 d'entre eux contiennent moins de cinq foyers sinistrés. Ces polygones sont représentés en gris clair, et ceux contenant au moins cinq observations sont représentés en gris foncé.

De plus, pour l'analyse des données surfaciques, nous avons aussi pris en compte certaines AD ontariennes liées aux foyers sinistrés enregistrés par la CRC en Ontario. En effet, dans la mesure où certaines zones d'eau (notamment la rivière des Outaouais) sont partagées en deux provinces, il est pertinent de ne pas se limiter aux frontières administratives qui peuvent altérer les résultats. Ainsi, 30 AD incluant au minimum cinq foyers sinistrés ontariens ont été inclues dans les données pour pouvoir établir une structure de voisinage plus réaliste.

Concernant l'absence de données du recensement par Statistique Canada, cette ano-

FIGURE 4.20 – Visualisation des aires de diffusion associées aux foyers sinistrés (Québec). *En gris clair : AD avec au moins un foyer sinistré. En gris foncé : AD avec au moins cinq foyers sinistrés.*

malie ne concerne qu'une AD avec suffisamment d'observations (cinq précisément) qui a donc été supprimée du jeu de données. La suite des analyses s'effectuera ainsi sur ces 177 AD (147 québécoises et 30 ontariennes) étant liées à au moins cinq observations. La figure 4.21 illustre ces 177 AD sélectionnées, avec en bleu celles appartenant à la province du Québec et en jaune celles de l'Ontario. De nouvelles variables ont été créées comme le montant moyen reçu par foyer (par AD), le nombre moyen de personnes par foyer (par AD), le total des bénéficiaires (par AD), le nombre de foyers sinistrés (par AD) ou encore le pourcentage des bénéficiaires par rapport à la population de 2016 (par AD). Par ailleurs, les catégories de la variable *sévérité* ont été transformées en valeurs numériques (1 pour mineure, 2 pour modérée, 3 pour importante et 4 pour extrême). Le tableau 4.8 récapitule les statistiques descriptives de toutes les variables disponibles.

4.3.2 Définition de la structure de voisinage

La première étape consiste à définir une structure de voisinage adaptée aux données de l'étude. Cette section présentera les implications liées aux différents choix possibles pour définir la matrice de voisinage des aires de diffusion, ainsi que les matrices de poids

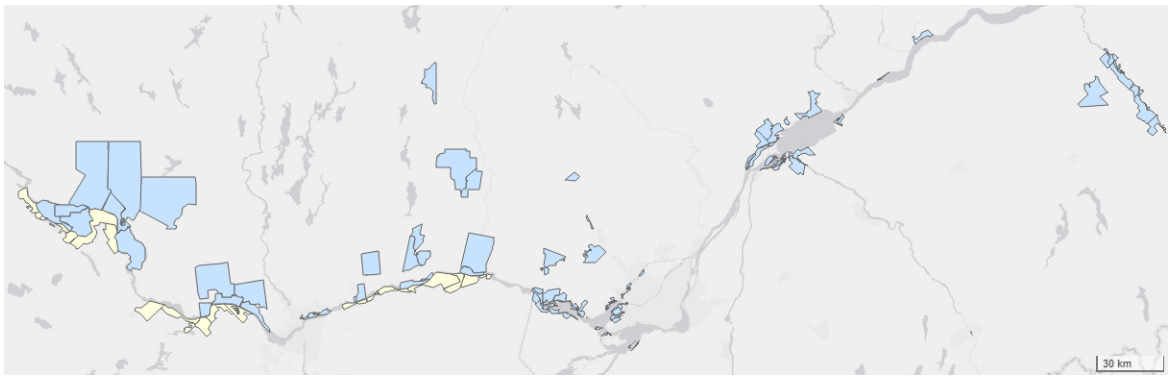


FIGURE 4.21 – Visualisation des aires de diffusion sélectionnées pour l'analyse des données surfaciques.
En bleu : AD québécoises. En jaune : AD ontariennes.

qui peuvent leur être associées.

Définition des voisins

Le tableau 4.9 et la figure 4.22 présentent les résultats des différentes méthodes de définition des voisins. On constate aisément que la méthode de définition du voisinage selon les voisins à la distance minimale n'est pas vraiment pertinente dans la mesure où la présence d'aires de diffusion isolées impose une distance minimale très élevée (près de 40 km). Comme mentionné à la section 3.1.2, on retrouve une grande disparité dans le nombre de voisins (Bivand et collab., 2013) avec, par exemple, 23 régions connectées chacune à 62 autres, alors que quatre régions n'ont qu'un seul voisin. Ainsi, le nombre moyen de liens est beaucoup plus élevé pour cette configuration que pour les autres, avec 33,1 voisins en moyenne pour chaque AD.

Selon certaines configurations, plusieurs AD n'ont pas de voisins (jusqu'à 50 AD pour la configuration voisins relatifs). Le nombre de polygones isolés est d'autant plus élevé que l'on a supprimé les AD n'ayant pas assez d'observations pour être intéressantes. Ce choix impacte aussi la distance moyenne entre les AD voisines pour les configurations des k voisins les plus proches.

Les matrices de voisinage issues de ces configurations seront converties en matrices

	Min	1er Q.	Med	Moy	3ème Q.	Max
Montant moyen	267,3	700,0	955,3	1 162,1	1 245,1	3 952,8
Total montants	1 336	6 813	14 400	45 864	39 000	615 566
NB moyen/p/foyer	1,1	1,8	2,1	2,1	2,4	3,5
Total bénéficiaires	6	15	29	72	78	867
NB foyers	5	8	13	32,8	35	330
% sexe féminin	11,1	46,7	50	50,4	53,8	88,9
% 0 à 5 ans	0	0	2,9	3,9	6,1	18,2
% 5 à 18 ans	0	4,6	11,1	11,3	16,2	40
% 18 à 65 ans	20,0	53,8	60,0	60,7	66,7	100
% + de 65 ans	0	12,4	18,5	21,2	27,6	80
Population 2016	259	483	589	730	720	5210
% Tot bénéf/Pop 2016	0,2	2,6	4,3	11,6	12,5	94,32
Âge moyen	30,7	41,2	43,7	44,1	46,7	69,2
Taille moy ménages	1,5	2,1	2,3	2,3	2,5	3,2
Revenu med ménage	20 064	50 880	61 312	64 898	77 184	134 656
Taux d'activité	33,9	55,7	62,7	62,2	68,9	81,9
Taux d'emploi	33,9	51,5	58,0	58,3	65,2	77,8
Taux de chômage	0	3,8	6,1	6,4	8,5	16,9
Valeur med. logements	100 192	180 217	219 387	242 046	279 284	602 107
Valeur moy. logements	108 830	194 022	243 376	274 785	318 441	786 614
	Mineure = 1	Modérée = 2	Importante = 3	Extrême = 4	Total	
Sévérité	8	17	42	110	177	

TABLEAU 4.8 – Statistiques descriptives pour les 177 aires de diffusion ayant au moins cinq observations. *Exemple : le montant moyen par foyer le plus faible pour une AD est de 267,30 $.*

de poids selon les différentes configurations et serviront pour la suite des analyses. Ces matrices de poids doivent pouvoir représenter au mieux l'intensité des relations géographiques entre les observations. Il est à noter que les matrices de voisinage utilisées dans les modèles spatiaux doivent respecter certains critères techniques et les zones sans voisins sont écartées.

4.3.3 Autocorrélations globale et locale

La définition d'une structure de voisinage pour les 177 aires de diffusion des données va pouvoir permettre de mesurer la dépendance spatiale et de répondre à certaines ques-

Définition du voisinage	Nombre de liens	Nb moyen de liens	Distance minimale	Distance moyenne	Distance maximale	Nombre d'AD sans voisins
Contiguïté Queen	440	2,5	179,2	4 747,4	23 165,4	22
Contiguïté Rook	416	2,3	179,2	4 772,1	23 165,4	23
Triang. de Delaunay	1030	5,8	179,2	18 917,4	218 842	0
Sphère d'influence	460	2,6	179,2	5 514,6	39 816,9	0
Graphe de Gabriel	253	1,4	179,2	10 745,6	152 804,2	41
Voisins relatifs	188	1,1	179,2	7 190,2	108 645,4	50
Voisin le plus proche	177	1,0	179,2	3 885,7	39 816,9	0
2 V. les plus proches	354	2,0	179,2	5 437,4	109 092,8	0
3 V. les plus proches	531	3,0	179,2	6 491,8	111 519,6	0
V. à la dist. minimale	5 856	33,1	179,2	18 829,0	39 816,9	0

TABLEAU 4.9 – Analyse descriptive des structures de voisinage pour les 177 aires de diffusion (AD) liées à au minimum cinq foyers sinistrés. *Note : Les distances sont exprimées en mètres.*

tions. Est-ce que le montant moyen versé pour un foyer d'une AD est lié aux montants versés pour les foyers de ses AD voisines ? En d'autres termes, peut-on considérer que la répartition spatiale du montant moyen versé par foyer est complètement aléatoire ?

Pour commencer à répondre à ces questions, une comparaison entre la distribution réelle et une distribution simulée (de manière complètement aléatoire) des montants moyens reçus par foyers ($ CAD) selon les AD a été effectuée. En raison d'enjeux de confidentialité, ces cartes ne sont pas publiées ici, cependant on observe à première vue une différence dans la répartition des valeurs. Par exemple, sur la carte de la distribution réelle les AD situées à l'Ouest semblent se regrouper selon la première ou la deuxième catégorie de montants moyens (entre 0$ et 2 000$) et de manière plus générale, les valeurs similaires apparaissent souvent côte-à-côte. Concernant la carte de la distribution simulée, on pourrait penser que la distribution aléatoire présente davantage de valeurs élevées, car elles sont beaucoup plus visibles que sur la carte de la distribution réelle. Cependant, elle a été simulée grâce à des permutations et les valeurs sont ainsi identiques à celle de la distribution réelle. La distribution simulée met alors en lumière des valeurs élevées qui n'apparaissent pas à une aussi grande échelle sur la carte réelle et qui sont regroupées dans des petites AD au niveau du Grand-Montréal par exemple. Le contraste entre ces deux cartes semble témoigner d'une dépendance entre les valeurs des AD voisines. L'étude des

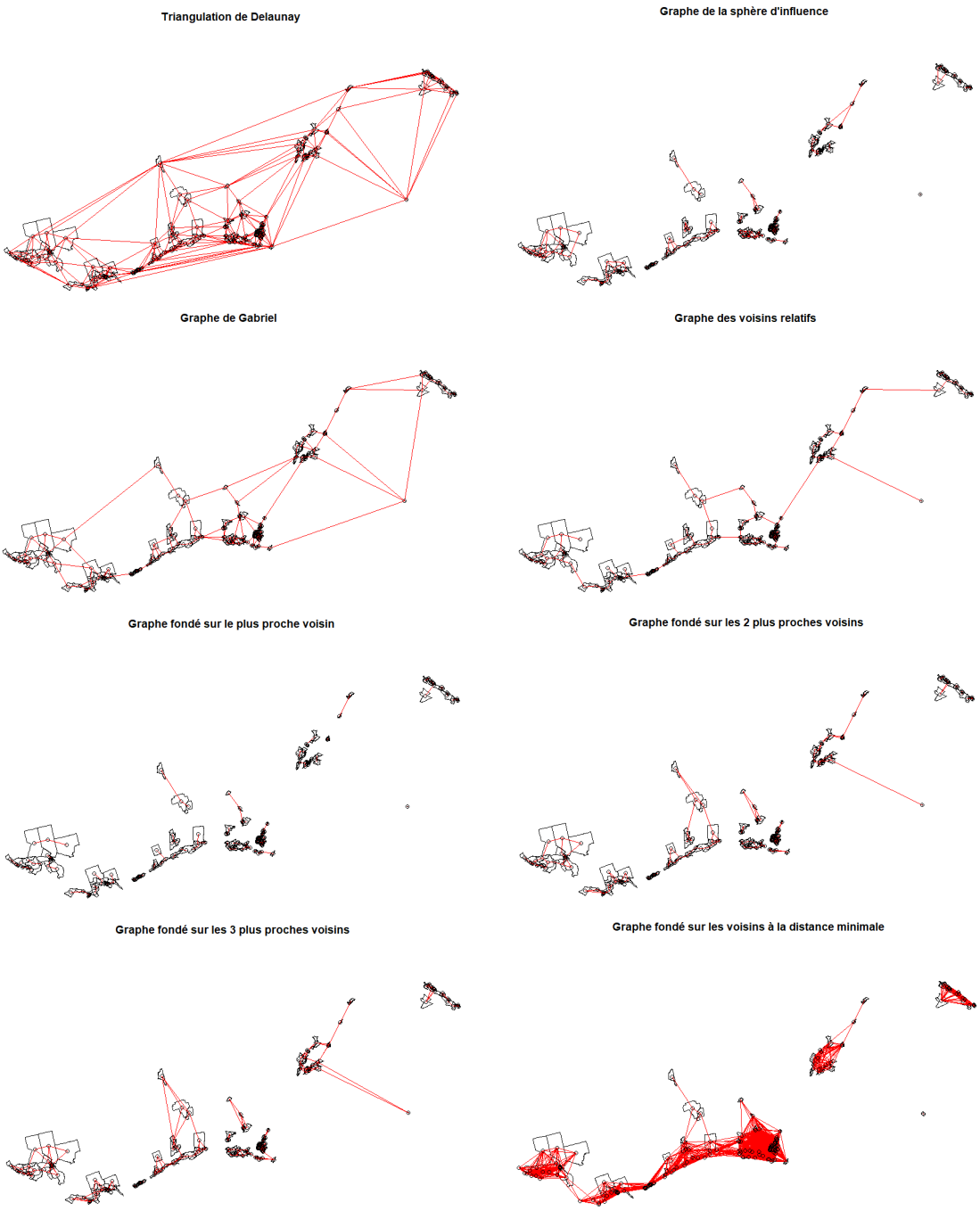
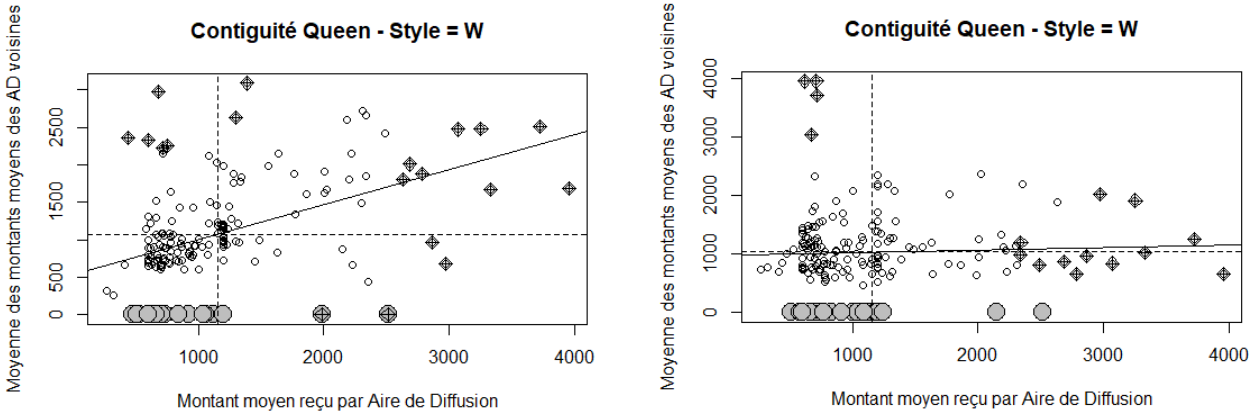


FIGURE 4.22 – Représentation des relations entre aires de diffusion selon la définition de la structure de voisinage.
Note : seules les AD liées à au moins cinq foyers sinistrés ont été prises en compte dans les analyses.

diagrammes de Moran et des indices d'autocorrélation ont alors pour but d'affirmer ou de contester cette hypothèse.

La différence entre ces deux cartes se confirme avec les diagrammes de Moran qui permettent d'avoir une idée de l'autocorrélation globale. La figure 4.23 oppose le diagramme de la distribution réelle (gauche) à celui de la distribution aléatoire présentée précédemment (droite). On remarque bien que la droite de régression pour le diagramme de droite a une pente proche de 0, alors que celle de gauche est positive. Cette pente de 0,38 correspond au I de Moran global, calculé selon la contiguïté Queen et avec une normalisation en ligne (W) de la matrice de poids. Sur les diagrammes, les petites croix indiquent les valeurs ayant une influence notable sur le I de Moran. La structure de voisinage basée sur la notion de contiguïté autorise les zones sans voisins et les cercles gris représentent les valeurs liées à ces AD isolées. Le nombre d'observations situées dans le cadrant « Low-Low » semble important et il serait intéressant de le vérifier avec l'indice de Moran local. Il est à noter que selon la structure de voisinage et le type de normalisation, les diagrammes de Moran ne sont pas tous similaires mais suivent cependant la même tendance.



(a) Distribution réelle (b) Distribution aléatoire simulée

FIGURE 4.23 – Diagrammes de Moran des montant moyens par AD pour la distribution réelle (gauche) et pour une distribution simulée aléatoirement avec permutations (droite).

Le tableau 5 de l'annexe 4.3.4 présente les différentes valeurs de l'indice de Moran global selon la structure de voisinage et le choix de normalisation de la structure de poids.

Définition du voisinage	I de Moran locaux			valeurs-p		
	Min	Moy	Max	Min	Moy	Max
Contiguïté Queen	-1,87	0,38	7,41	0	0,38	0,97
Triang. de Delaunay	-1,10	0,38	6,19	0	0,34	0,99
Sphère d'influence	-1,17	0,38	6,19	0	0,39	0,97
Graphe de Gabriel	-2,66	0,28	6,34	0	0,40	0,99
Voisins relatifs	-2,66	0,22	6,25	0	0,42	0,99
Voisin le plus proche	-1,88	0,28	10,52	0	0,43	0,97
2 V. les plus proches	-1,54	0,34	5,89	0	0,40	0,99
3 V. les plus proches	-1,39	0,34	7,06	0	0,39	0,99
V. à la distance min	-0,54	0,23	2,77	0	0,36	0,99

TABLEAU 4.10 – Statistiques descriptives des indices de Moran locaux et de leur valeurs-p. *Note : Utilisation des matrices de poids une avec normalisation en ligne.*

Pour les 33 tests, l'hypothèse nulle H_0 est rejetée, ce qui indique la présence d'une autocorrélation spatiale au sein des données. Selon la configuration et la normalisation, les indices de Moran globaux vont de 0,15 (configuration des voisins à distance minimale) à 0,46 (contiguïté Queen). On observe ainsi que lorsque le nombre de liens entre les AD augmente, l'indice de Moran diminue et on retrouve des écarts importants selon la définition de la structure de voisinage. On peut aussi noter que les écarts selon les types de normalisation restent relativement faibles.

Concernant l'autocorrélation locale, le tableau 4.10 résume la distribution des I de Moran locaux pour l'ensemble des AD, ainsi que celle des valeurs-p pour les différentes structures (avec normalisation en ligne). On observe ainsi que selon toutes les structures de voisinage, il y a une présence d'autocorrélation locale négative (valeurs minimums) qui contrastent avec des plus fortes valeurs du I de Moran local allant, par exemple, jusqu'à 10,58 pour la configuration du *voisin le plus proche*. Cependant, les valeurs des valeurs-p sont relativement élevée, avec des moyennes bien supérieures à 0,05 ce qui indique que de nombreuses valeurs du I de Moran local ne sont pas significatives.

Pour aller plus loin, les cartes de la figure 4.24 représentent les AD selon les résultats liés aux indices de Moran locaux. En violet, ce sont les 132 AD avec une valeur-p non significative et en gris il s'agit des 22 AD isolées (sans voisines). Parmi les AD avec des

Méthodes d'ajustement	Min	1er Q.	Med	Moy	3ème Q.	Max	Nb d'indices significatifs
Sans ajust.	*0*	*0.25*	*0.41*	*0.38*	*0.50*	*0.97*	*23*
Holm	0	1	1	0,9	1	1	13
Hochberg	0	0,97	0,97	0,88	0,97	0,97	13
BH	0	0,61	0,61	0,56	0,61	0,97	16
Bonferroni	0	1	1	0,9	1	1	13
FDR	0	0,61	0,61	0,56	0,61	0,97	16

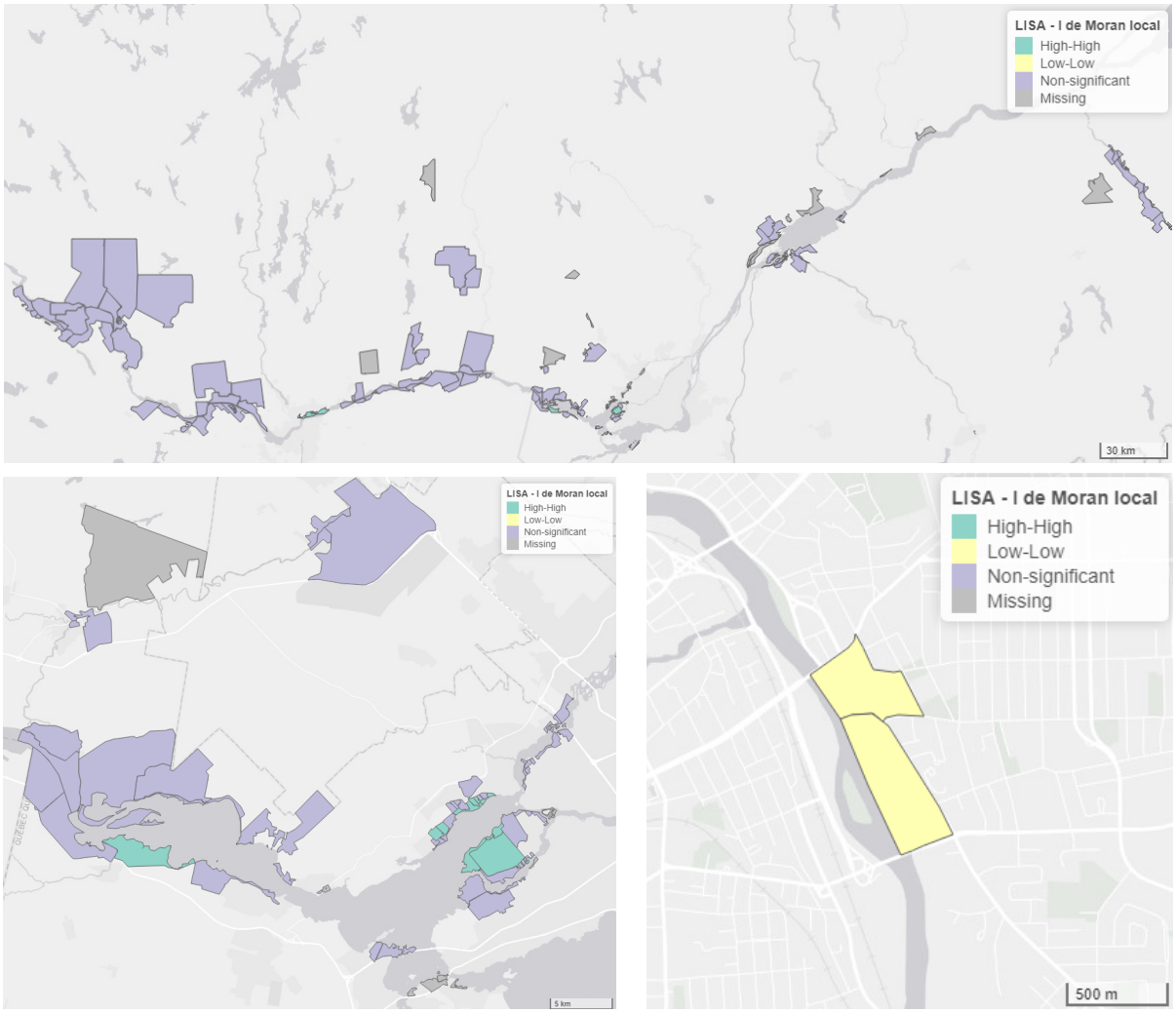
TABLEAU 4.11 – Distribution des valeurs-p des indices de Moran locaux, selon la méthode d'ajustement utilisée. *Structure de voisinage basée sur la contiguité Queen avec normalisation en ligne.*

valeurs-p significatives, il y en a 21 avec une structure « High-High » (montant élevé dans un environnement élevé) et uniquement deux avec une structure « Low-Low » (montant faible dans un environnement faible). Représentées en jaune sur la carte (b), ces deux petites AD sont situées à Sherbrooke et les valeurs moyennes versées par foyer sont de 323$ et 267$ respectivement. Ces valeurs correspondent bien à une structure « Low-Low » : valeurs de la variable plus faibles que la moyenne et situées dans un voisinage qui leur ressemble.

Cependant, comme nous l'avons vu au chapitre 3, il est important de faire preuve de prudence quant à l'interprétation des valeurs de significativité. Le tableau 4.11 montre en effet qu'en ajustant les valeurs-p pour tenir contre de l'augmentation du risque de détection erronée d'une autocorrélation, les valeurs-p sont beaucoup plus élevées. Par conséquent, le nombre d'AD ayant un indice d'autocorrélation spatiale local significatif diminue (environ une dizaine d'AD en moins). Ainsi, sur 177 AD, il n'y en a finalement que très peu qui présentent une autocorrélation significative (majoritairement positive).

4.3.4 Modélisation

L'analyse de l'autocorrélation via les indices global et local de Moran a mis en évidence l'existence d'une autocorrélation positive entre les valeurs du montant moyen alloué par foyer sinistré. En gardant à l'esprit une certaine prudence quand à l'existence de



(a) *Grand Montréal* (b) *Sherbrook*

FIGURE 4.24 – Significativité et structure dominante selon l'indice de Moran local.

réelles interactions spatiales, nous allons tenter de modéliser cette autocorrélation à l'aide des modèles d'économétrie spatiale présentés au chapitre 3.

En premier, un modèle de régression linéaire classique (non-spatial) est estimé avec la méthode des MCO. En comparant les résultats de plusieurs modèles de RLC, tous avec une sélection différente de variables, nous retenons un modèle à dix variables obtenant le meilleur AIC. Les résultats de ce modèle sont présentés au tableau 4.13 (1ère colonne : RLC). Nous garderons ces dix variables pour construire les prochains modèles spatiaux. Les corrélations entre ces variables et notre variable d'intérêt, le montant moyen versé par

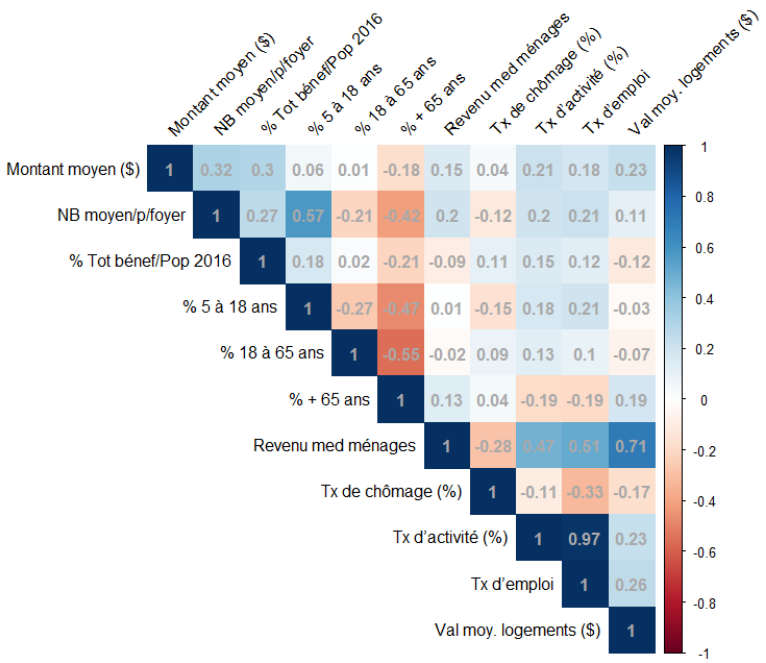


FIGURE 4.25 – Corrélations entre les 10 variables explicatives sélectionnées et la variable d'intérêt *Montant moyen*.

foyer selon les AD, sont disponibles à la figure 4.25.

À titre d'exemple, le AIC du modèle n'utilisant que les variables issues des données de la CRC est de 2 790, celui utilisant le maximum de variables disponibles (en prenant soin d'enlever certaines variables pour ne pas avoir de combinaison linéaire parfaite) est de 2 780, et celui de la meilleure sélection de variables est de 2 772. Il est aussi important de noter que l'incorporation des variables externes (Statistique Canada) permet d'augmenter le R^2 ajusté du modèle (passant de 0,15 à 0,26 pour la sélection des 10 variables les plus pertinentes).

En effectuant un test de Moran adapté sur les résidus du modèle pour chaque matrice de voisinage, on remarque que les résultats confirment la présence résiduelle d'une autocorrélation spatiale significative uniquement pour certaines d'entre-elles (voir tableau 4.12). En effet, les valeurs-p des matrices des voisins relatifs et du voisin le plus proche ne sont pas significatives (< 0.05).

Ensuite, les modèles spatiaux SEM, SAR, SDM et SLX sont estimés pour chacune

	Contiguïté	Delaunay	Influence	Gabriel	V. Relatifs	1 NN	2 NN	3 NN	Dist. min
I de Moran	0,22	0,20	0,18	0,15	0,10	0,08	0,17	0,15	0,09
valeur-p	0,0007	3,5e-07	0,002	0,007	0,06	0,13	0,002	0,0009	5,83-06

TABLEAU 4.12 – Résultats du test de Moran adapté sur les résidus du modèle RLC selon la matrice de voisinage (normalisation en ligne).

des matrices de voisinage. On présente au tableau 4.13 les résultats de ces différentes modélisations pour la matrice de contiguïté, car c'est celle démontrant généralement les meilleurs AIC (mêmes si ils restent tous assez similaires). Si cette matrice présente le caractère explicatif le plus élevé, elle est aussi une des plus simple à interpréter. Par ailleurs, comme mentionné à la section 3.1.1, il est souvent plus pertinent d'utiliser une matrice de contiguïté lorsque l'on traite de données démographiques et sociales car la séparation par une frontière administrative peut avoir plus d'importance qu'une mesure de distance entre les centroïdes.

Concernant le choix du modèle, les figures 4.26, 4.27 et 4.28 présentent respectivement les approches de décision ascendante, descendante et mixte. La première, qui permet de choisir entre une RLC, un SAR ou un SEM, favorise l'intégration d'une variable endogène spatialement décalée (modèle SAR) avec $\rho = 0,33$. Ainsi, avec ce modèle, le montant d'une AD i sera impactée positivement par les montant des AD possédant au moins un point de frontière en commun. L'approche descendante de LeSage et Pace (2009) ne semble pas adaptée aux données, car en partant d'un SDM, les tests du rapport de vraisemblance (LR_θ, LR_ρ) et le test du facteur commun (LR_λ) sont tous significatifs, ce qui ne nous permet pas de délibérer ensuite entre une RLC, un SAR, un SLX ou un SEM. Pour l'approche séquentielle d'Elhorst (2010) (voir figure 4.28), c'est le modèle spatial de Durbin (SDM) qui est sélectionné et dont le AIC est de 2 755. On peut cependant remarquer que les valeurs-p des coefficients β ne sont pas significatives, excepté pour la variable du revenu médian des ménages.

Le tableau 4.14 présente les impacts directs, indirects et totaux pour les variables présentes dans le modèle SDM sélectionné. Sans surprise, l'impact direct du nombre moyen de personnes par foyer est positif et plutôt élevé. En revanche, certains résultats peuvent

	RLC	SEM	SAR	SDM	SLX
(Constante)	1 584 *	1 822	1 392,3 *	1 538 **	1 462 *
β *NB moyen/p/foyer*	264,6 *	200,0	185,8	219,4	241,7*
β *% Tot bénef/Pop 2016*	9,72 ***	9,16 ***	6,6 **	8,14***	8,15 **
β *% 5 à 18 ans*	−27,3 ***	−24,0 ***	−24,0 ***	−22,7 ***	-24,1 ***
β *% 18 à 65 ans*	−15,8 **	−16,4 **	−16,0 **	−15,5 **	-15,6 **
β *% + de 65 ans*	−20,7 ***	−21,2 ***	−18,3 ***	−20,3 ***	−20,4 ***
β *Revenu med ménages*	$-0,003$	$-0,005$	$-0,004$	0,007 *	0,005
β *Taux de chômage*	−63,7 *	−60,0 *	−57,1 *	−68,3 **	−70,8 **
β *Taux d'activité*	132,8 **	114,8 ***	112,5 **	129,6 ***	137,2 ***
β *Taux d'emploi*	−129,8 **	−117,4 **	−110,9 **	−135,2 ***	−140,9 ***
β *Val moy. logements ($)*	0,002 ***	0,001 *	0,001 *	0,001 **	0,001 **
ρ			0,42 ***	0,25 ***	
λ		0,33 ***			
θ *NB moyen/p/foyer*				−31,7	89,1
θ *Tot bénef/Pop 2016*				−1,8	0,4
θ *% 5 à 18 ans*				−5,2	−10,5
θ *% 18 à 65 ans*				2,2	0,4
θ *% + de 65 ans*				5,2	3,4
θ *Revenu med mén.*				−0,02 ***	−0,02 ***
θ *Taux de chômage*				−65,3	−73,0
θ *Taux d'activité*				152,1	173,9
θ *Taux d'emploi*				−141,0	−159,4
θ *Val moy. logt. ($)*				0,001	0,001 **
AIC	2 772	2 761	2 764	2755	2763
R^2 adj	0,26				0,33
Test Moran	0,22 (<0,001)				0,19 (0,002)
Test LM-Error (valeur-p)	0,001				0,004
Test LM-Lag (valeur-p)	<0,001				<0,001
Test Robuste LM-Error (valeur-p)	0,7997				0,08
Test Robuste LM-Lag (valeur-p)	0,0127				0,004
Test facteur commun (valeur-p)				0,003	

TABLEAU 4.13 – Déterminants du montant moyen alloué à chaque foyer par aire de diffusion, à partir d'une matrice basée sur la notion contiguïté (Queen) avec normalisation en ligne (W). Les tests LM correspondent aux tests du multiplicateur de Lagrange, définis à la section 3.3.2.
*Significativité : *** : p < 0,01, ** : p < 0,05, * : p < 0,1.*

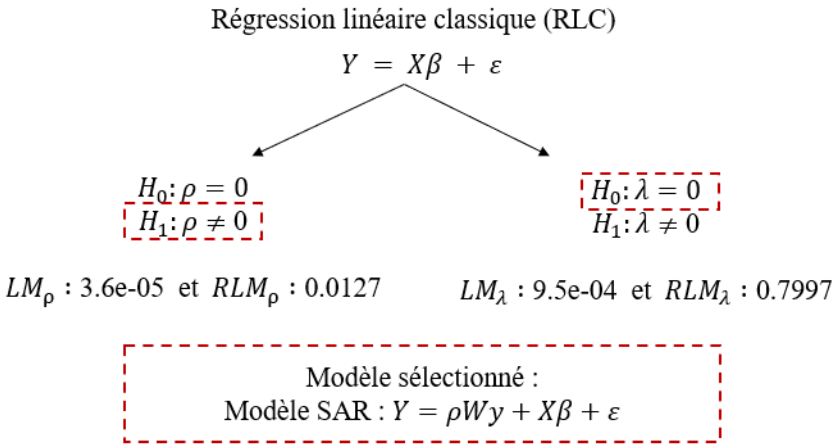


FIGURE 4.26 – Tests d'hypothèses pour comparer les modèles de régression spatiaux selon la méthode ascendante
Note : pour les tests, ce sont les valeurs-p qui sont affichées.

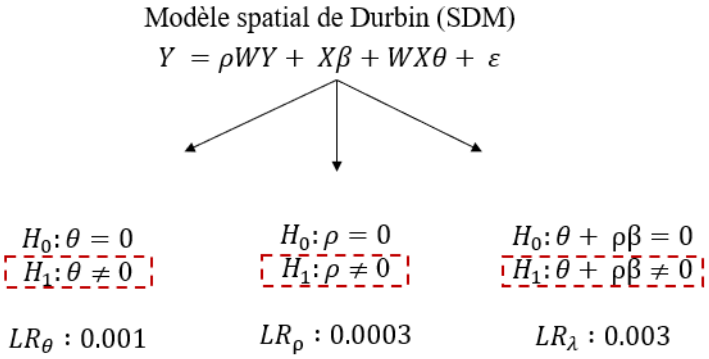


FIGURE 4.27 – Tests d'hypothèses pour comparer les modèles de régression spatiaux selon la méthode descendante.
Note : pour les tests, ce sont les valeurs-p qui sont affichées.

paraître étonnants, comme ceux liés aux taux de chômage, d'activité et d'emploi. Le taux de chômage et le taux d'emploi ont tout les deux des impacts négatifs assez importants, alors que le taux d'activité présente un effet positif. Cela s'explique par la forte corrélation entre le taux d'emploi et le taux d'activité.

Cependant, beaucoup d'incertitudes demeurent quand à l'interprétation des résultats

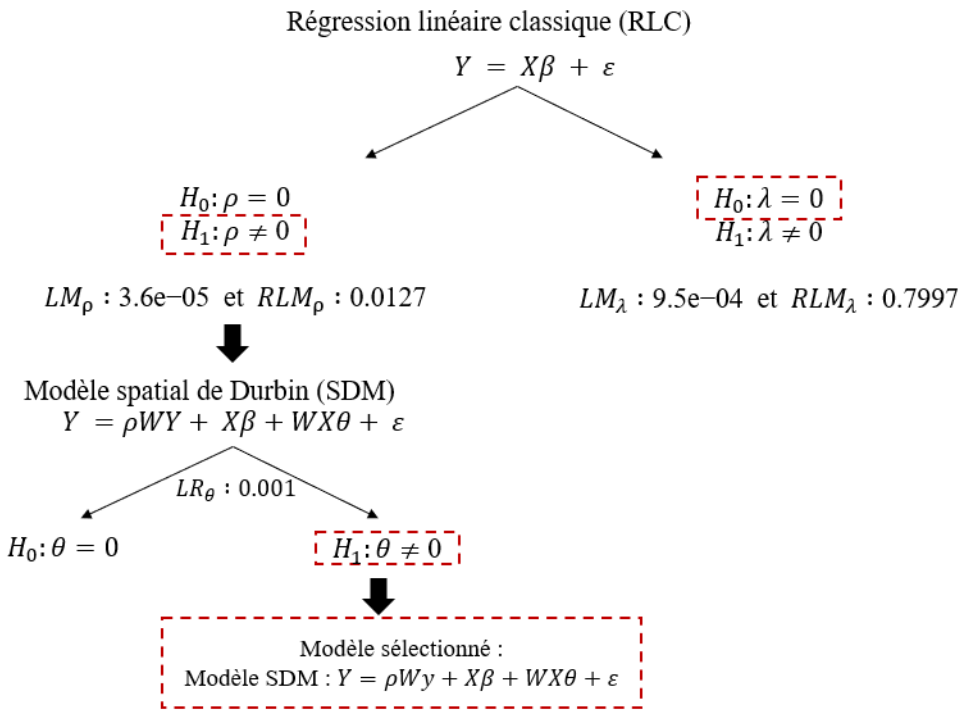


FIGURE 4.28 – Tests d'hypothèses pour comparer les modèles de régression spatiaux selon la méthode mixte.
Note : pour les tests, ce sont les valeurs-p qui sont affichées.

dans le contexte de recherche. Par exemple, le ratio du nombre de bénéficiaires par rapport à la population totale de l'AD a un impact positif sur le montant moyen, mais cela pourrait provenir de plusieurs aspects. On pourrait notamment penser que si une grande partie de la population d'une AD a reçu une aide financière de la CRC, cela peut signifier que les inondations ont été très intenses dans cette zone. Le montant moyen aura donc tendance à être plus élevé. Une seconde interprétation pourrait concerner les ressources financières et matérielles des ménages avant les sinistres. En effet, dans la mesure où toutes les victimes des inondations n'ont pas nécessairement sollicité l'aide de la CRC, il se pourrait qu'un faible ratio soit lié à l'absence d'un besoin d'aide financière. Même si elle reste faible, on observe, par exemple, une corrélation négative (figure 4.25) entre ce ratio et la valeur moyenne des logements.

Pour conclure, il faut rester prudents quant à l'interprétation des résultats, dans la me-

Variables	Impacts directs	Impacts indirects	Impacts totaux
NB moyen/p/foyer	221,5	27,2	248,7
% Tot bénef/Pop 2016	8,2	0,2	8,4
% 5 à 18 ans	-23,8	-13,3	37,1
% 18 à 65 ans	-15,6	-2,0	-17,6
% + de 65 ans	-20,3	0,4	-19,9
Revenu med ménages	0,005	-0,022	-0,017
Taux de chômage	-76,2	-100,9	-177,1
Taux d'activité	147,2	226,1	373,3
Taux d'emploi	-151,8	-214,2	-366,0
Val moy. logements ($)	0,001	0,002	0,002

TABLEAU 4.14 – Impacts directs, indirects et totaux du modèle SDM (matrice de contiguïté avec normalisation en ligne).

sure où beaucoup d'aspects incertains sont à prendre en compte. En effet, on peut par exemple citer l'absence d'information fiable concernant l'intensité des inondations dans les AD, qui est pourtant un aspect crucial lors de l'allocation des dons après un sinistre. La création d'une variable de sévérité pour les AD à partir des événements d'inondation de *Données Québec* s'est révélée non pertinente dans la sélection des variables. Ce résultat n'est en réalité pas très surprenant, car la méthode d'assignation des facteurs de sévérité aux AD depuis les localisations relevées par *DQ* est relativement imprécise. De plus, on peut questionner la pertinence de l'utilisation des AD comme unités spatiales. Comme mentionné au chapitre 1, les résultats peuvent être influencés par un effet de zonage car les frontières des AD sont souvent irrégulières et la taille des surfaces varie grandement. Par ailleurs, du fait de leur forme et de leur position dans l'espace, certaines AD sinistrées possèdent plusieurs kilomètres de rivage, alors que d'autres sont beaucoup moins exposées à des zones d'eau (rivières, fleuves, lacs). Ainsi, le découpage administratif ne correspond pas forcément à la réalité du versement des aides d'urgence suite à une catastrophe naturelle.

Conclusion

Lorsque des données possèdent des attributs liés à leur localisation, il est souvent pertinent de les prendre en compte et d'effectuer des analyses spécifiques à ce type de données. Ces informations, potentiellement riches pour l'analyse, permettent notamment de répondre à des questions qui n'appartiennent pas au cadre des méthodes d'analyse de données classiques. Toutes les données possédant une information de localisation ne se traitent pas de la même manière selon la nature des observations et la question de recherche associée. Si l'on distingue généralement trois types de données (Cressie, 1993), ceux-ci peuvent néanmoins être perméables. Par exemple, il est possible d'analyser la même information initiale de différentes manières, comme nous l'avons fait avec les données récoltées par la Croix-Rouge canadienne. En effet, à partir des attributs de chaque foyer sinistré et de ses coordonnées géographiques, nous avons pu utiliser à la fois des techniques conçues pour l'analyse des configurations de points, mais aussi pour l'analyse des données surfaciques. Pour ce faire, il était nécessaire de présenter au préalable les éléments théoriques permettant d'analyser les données relatives à l'aide financière apportée aux victimes suite aux inondations dans la province du Québec. Les notions et concepts fondamentaux liés aux données spatiales ont tout d'abord été introduits au chapitre 1 afin notamment de présenter les éléments essentiels à prendre en compte dans les analyses. Par la suite, le chapitre 2 a permis d'exposer différentes méthodes exploratoires de l'intensité et de la dépendance d'une configuration de points, avec une emphase particulière sur l'étude des processus marqués. La régression géographiquement pondérée (RGP) a été présentée comme technique de modélisation des données ponctuelles. Les méthodes

d'analyse des données surfaciques ont ensuite été développées au troisième chapitre, avec la définition d'une structure de voisinage et la détection puis la modélisation de l'autocorrélation spatiale. Ces trois premiers chapitres théoriques ont ainsi permis de présenter une sélection de méthodes d'analyse pouvant être appliquées de manière pertinente aux données de la CRC dans le chapitre 4.

Plusieurs aspects se sont révélés intéressants dans le cadre de cette étude de cas. Premièrement, la prise en compte de données externes s'est avérée très pertinente. En effet, l'analyse des données surfaciques a pu être réalisée grâce à l'utilisation d'un fichier externe définissant les unités géographiques sur lesquelles s'appuyer. Les données récoltées par la CRC ont ainsi pu être agrégées au niveau des aires de diffusion (AD), qui représentent les plus petites régions géographiquement normalisées pour lesquelles les données de recensement sont disponibles. Par ailleurs, outre la définition de ces aires à étudier, l'utilisation des données externes a aussi permis d'apporter des éléments supplémentaires, notamment démographiques comme les données de recensement de Statistique Canada. Si les modélisations effectuées à la fois sur les données ponctuelles et sur les données surfaciques ont augmenté les critères de performance des modèles, ces améliorations restent cependant légères. L'utilisation de ces variables supplémentaires semblent alors démontrer l'absence d'impact des facteurs socio-économiques et démographiques sur la demande d'aide des foyers sinistrés à la CRC.

Deuxièmement, l'utilisation de l'information spatiale lors de la modélisation de l'allocation des dons aux sinistrés s'est avérée judicieuse et utile. Concernant l'analyse des données ponctuelles, les différents modèles de RGP ont en effet démontré une meilleure performance que les modèles de régression classique, permettant ainsi de mettre en valeur l'importance de la localisation des foyers sinistrés dans la modélisation des aides versées par la CRC. Par ailleurs, au travers des modèles d'économétrie spatiale, l'analyse surfacique confirme l'intérêt d'utiliser l'information spatiale disponible. L'approche d'Elhorst (2010) conduit ainsi à la sélection du modèle spatial de Durbin qui définit des interactions endogènes et exogènes. En d'autres termes, les valeurs des régions voisines semblent s'influencer entre-elles et le montant moyen versé d'une AD apparaît alors comme dépendant

de ses propres caractéristiques mais aussi de celles de ses voisines.

Pour finir, cette étude de cas a permis de faire ressortir des anomalies qui ont été détectées parmi les données de la CRC. Par exemple, si la qualité générale des coordonnées géographiques récoltées est très bonne, on peut cependant citer les superpositions de certaines coordonnées qui, dans le contexte d'étude, ne semblent pas être fidèles à la réalité. La mise en lumière de ces aspects au sein des bases de données peut alors permettre à la CRC d'identifier différentes pistes d'amélioration quant à leur système de collecte et de stockage des données. Il est néanmoins nécessaire de rappeler l'aspect particulier et complexe de la collecte de données en situation de crise humanitaire. En opérant dans un contexte d'urgence où la priorité est d'apporter de l'aide aux personnes sinistrés, la collecte de données doit se faire rapidement pour ne pas compromettre l'objectif principal.

Certaines limites nécessitent cependant d'être mentionnées car outre l'identification d'anomalies, différents aspects intrinsèques aux données compliquent l'interprétation des résultats et ne permettent pas d'établir des conclusions assurément fiables. Par exemple, l'analyse descriptive a révélé une faible variabilité du montant reçu par foyer ($ CAD), avec notamment 45% des foyers ayant reçu une aide totale de 600$. Cela s'explique par le déploiement du programme d'assistance financière à grande échelle qui vise à rapidement aider les foyers impactés à couvrir les dépenses liées à l'évacuation d'urgence en leur versant une somme unique. De plus, si l'intégration de l'information spatiale et des données externes permet d'augmenter les performances des modèles de régression (RLC vs. RGP), la variabilité du montant alloué reste mal expliquée par les données. Ainsi, dans le cadre de la RGP, les données disponibles, qu'elles soient internes ou externes, ne permettent d'expliquer qu'une petite partie (moins de 20%) de l'allocation financière attribuée aux foyers sinistrés. Concernant la modélisation du montant moyen par aire de diffusion, les résultats proposés sont aussi à interpréter avec précaution. Si l'approche d'Elhorst (2010) privilégie un SDM impliquant des interactions endogènes et exogènes, les coefficients relatifs aux effets d'interactions exogènes ne sont cependant pas significatifs. Enfin, les tests des I de Moran locaux ont démontré que seulement une minorité d'AD présentaient une autocorrélation significative.

Enfin, certaines limites permettent de mettre en avant un important potentiel d'amélioration pour de futures modélisations de l'allocation des dons suite à des catastrophes naturelles. En effet, plusieurs variables présentes dans la base de données initiale ont été écartées alors que leurs capacités explicatives semblaient pertinentes dans le contexte de l'étude. Des informations relatives aux revenus des foyers, à leurs couvertures d'assurance ou encore à la gravité des dommages occasionnés gagneraient à être collectées afin d'améliorer les performances de modélisation du montant reçu. Ainsi, qu'il s'agisse de données classiques ou bien spatiales, une bonne collecte de données fait partie des principaux défis à relever afin de pouvoir dégager des conclusions fiables et utiles pour le déploiement des prochaines opérations de soutien financier suite à une catastrophe naturelle. Toutefois, il est important de considérer le compromis à effectuer entre la quantité d'information récoltée pour chaque foyer et le temps de déploiement de certains programme comme celui de l'assistance à grande échelle qui vise à aider le plus de sinistrés le plus rapidement possible.

En définitive, cette étude de cas a permis d'illustrer l'important potentiel de valorisation des données récoltées par la CRC et de leur information spatiale. Elles peuvent à la fois s'analyser au niveau des foyers individuels avec l'analyse des données ponctuelles, mais aussi à un niveau d'agrégation plus élevé grâce à la transformation de l'unité de référence des données. Le niveau d'agrégation privilégié correspond aux aires de diffusion qui, contrairement aux codes postaux, sont délimitées et publiées de manière fiable par le gouvernement canadien. Par ailleurs, toutes les analyses semblent confirmer que le nombre de personnes présentes dans un foyer sinistré est un facteur ayant un impact significatif sur l'allocation des dons aux foyers sinistrés.

Pour finir, avec la multiplication des événements climatiques extrêmes engendrés par le dérèglement climatique (IPCC, 2018), le déploiement de bonnes collectes et de valorisation des données va devenir de plus en plus crucial. En effet, selon l'Agence de la santé publique du Canada (2021), la configuration variable des pluies, les tempêtes plus violentes, la fonte rapide des neiges et l'élévation du niveau de la mer en raison des changements climatiques vont accroître le risque d'inondation au Canada. Devant ce défi,

l'exploitation des outils et ressources techniques offertes par la science de données apparaît alors comme essentielle pour la mise en place d'opérations de soutien aux populations face aux futures catastrophes.

Bibliographie

Agence de la santé publique du Canada. 2021, «Changements climatiques, inondations et votre santé», URL `https://www.canada.ca/fr/sante-publique/services/promotion-sante/sante-publique-environnementale-changements-climatiques/fiches-information-changements-climatiques-sante-publique-inondations.html`.

Anselin, L. 1988, *Spatial Econometrics : Methods and Models*, Kluwer Academic Publishers.

Anselin, L. 1995, «Local indicators of spatial association—LISA», *Geographical analysis*, vol. 27, n° 2, p. 93–115.

Anselin, L., A. K. Bera, R. J. G. M. Florax et M. J. Yoon. 1996, «Simple diagnostic tests for spatial dependence», *Regional science and urban economics*, vol. 26, n° 1, p. 77–104.

Anselin, L. et R. J. G. M. Florax. 1995, «Small sample properties of tests for spatial dependence in regression models : Some further results», dans *New Directions in Spatial Econometrics*, édité par L. Anselin et R. J. G. M. Florax, Springer, p. 21–74.

Anselin, L. et D. A. Griffith. 1988, «Do spatial effecfs really matter in regression analysis ?», *Papers in Regional Science*, vol. 65, n° 1, p. 11–34.

Anselin, L. et S. Rey. 1991, «Properties of tests for spatial dependence in linear regression models», *Geographical analysis*, vol. 23, n° 2, p. 112–131.

Avis, D. et J. Horton. 1985, «Remarks on the sphere of influence graph», *Annals of the New York Academy of Sciences*, vol. 440, n° 1, p. 323–327.

Baddeley, A., E. Rubak et R. Turner. 2015, *Spatial Point Patterns : Methodology and Applications with R*, CRC press.

Baddeley, A. J., J. Møller et R. Waagepetersen. 2000, «Non-and semi-parametric estimation of interaction in inhomogeneous point patterns», *Statistica Neerlandica*, vol. 54, n° 3, p. 329–350.

Bavaud, F. 1998, «Models for spatial weights : a systematic look», *Geographical analysis*, vol. 30, n° 2, p. 153–171.

Besag, J. 1977, «Comments on Ripley's paper», *Journal of the Royal Statistical Society*, vol. 39, p. 193–195.

Bivand, R. S., E. Pebesma et V. Gómez-Rubio. 2013, «Spatial neighbors», dans *Applied spatial data analysis with R*, Springer, p. 83–125.

Bouayad Agha, S. et M.-P. De Bellefon. 2018, «Indices d'autocorrélation spatiales», dans *Manuel d'analyse spatiale. Théorie et mise en œuvre pratique avec R*, Institut National de la Statistique et des Études Économiques.

Brunsdon, C., A. S. Fotheringham et M. E. Charlton. 1996, «Geographically weighted regression : a method for exploring spatial nonstationarity», *Geographical analysis*, vol. 28, n° 4, p. 281–298.

Brunsdon, C., S. Fotheringham et M. Charlton. 1998, «Geographically weighted regression», *Journal of the Royal Statistical Society : Series D (The Statistician)*, vol. 47, n° 3, p. 431–443.

Cheong, F. et C. Cheong. 2011, «Social media data mining : A social network analysis of tweets during the 2010–2011 australian floods.», *15th Pacific Asia Conference on Information Systems (PACIS)*, vol. 11, p. 46–46.

Chiles, J.-P. et P. Delfiner. 2009, *Geostatistics : Modeling Spatial Uncertainty*, vol. 497, John Wiley & Sons.

Cressie, N. A. 1993, *Statistics for spatial data*, John Wiley & Sons.

De Bellefon, M.-P., V. Loonis et R. Le Gleut. 2018, «Codifier la structure de voisinage», dans *Manuel d'analyse spatiale. Théorie et mise en œuvre pratique avec R*, Institut National de la Statistique et des Études Économiques.

Diggle, P. 1985, «A kernel method for smoothing point process data», *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, vol. 34, n° 2, p. 138–147.

Diggle, P. J. et A. G. Chetwynd. 1991, «Second-order analysis of spatial clustering for inhomogeneous populations», *Biometrics*, p. 1155–1163.

Données Québec. 2019, «Fichier des événements d'inondation - printemps 2019», URL `https://www.donneesquebec.ca/recherche/dataset/cartographie-des-inondations-printemps-2019/resource/603fda9c-54a2-4099-baf8-bdd5135af712?inner_span=True`.

Dontas, E., F. Toufexis, N. Bardis et N. Doukas. 2017, «Data acquisition for environmental and humanitarian crisis management», dans *Green IT Engineering : Components, Networks and Systems Implementation*, Springer, p. 87–107.

Elhorst, J. P. 2010, «Applied spatial econometrics : raising the bar», *Spatial Economic Analysis*, vol. 5, n° 1, p. 9–28.

Ellison, G. et E. L. Glaeser. 1997, «Geographic concentration in US manufacturing industries : a dartboard approach», *Journal of Political Economy*, vol. 105, n° 5, p. 889–927.

Floch, J.-M. et R. Le Saout. 2018, «Économétrie spatiale : modèles courants», dans *Manuel d'analyse spatiale. Théorie et mise en œuvre pratique avec R*, Institut National de la Statistique et des Études Économiques.

Floch, J.-M., E. Marcon et F. Puech. 2018, «Les confugurations de points», dans *Manuel d'analyse spatiale. Théorie et mise en œuvre pratique avec R*, Institut National de la Statistique et des Études Économiques.

Florax, R. J. G. M. et H. Folmer. 1992, «Specification and estimation of spatial linear regression models : Monte Carlo evaluation of pre-test estimators», *Regional Science and Urban Economics*, vol. 22, n° 3, p. 405–432.

Florax, R. J. G. M., H. Folmer et S. J. Rey. 2003, «Specification searches in spatial econometrics : the relevance of Hendry's methodology», *Regional Science and Urban Economics*, vol. 33, n° 5, p. 557–579.

Fotheringham, A. S., C. Brunsdon et M. Charlton. 2003, *Geographically Weighted Regression : the analysis of spatially varying relationships*, John Wiley & Sons.

Geary, R. C. 1954, «The contiguity ratio and statistical mapping», *The Incorporated Statistician*, vol. 5, n° 3, p. 115–146.

Gollini, I., B. Lu, M. Charlton, C. Brunsdon et P. Harris. 2015, «Gwmodel : an R package for exploring spatial heterogeneity using geographically weighted models», *Journal of Statistical Software*, vol. 63, n° 17, p. 1–50.

Harkness, R. et V. Isham. 1983, «A bivariate spatial point pattern of ants' nests», *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, vol. 32, n° 3, p. 293–303.

Holm, S. 1979, «A simple sequentially rejective multiple test procedure», *Scandinavian Journal of Statistics*, p. 65–70.

IPCC. 2018, «Summary for policymakers.», *Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and*

related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty.

Kelejian, H. H. et I. R. Prucha. 2010, «Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances», *Journal of Econometrics*, vol. 157, n° 1, p. 53–67.

Kooijman, S. 1979, «The description of point patterns», *Spatial and Temporal Analysis in Ecology*, p. 305–331.

Le Gallo, J. 2000, *Econométrie spatiale (1, Autocorrélation spatiale)*, thèse de doctorat, Laboratoire d'analyse et de techniques économiques (LATEC), Université de Bourgogne.

Le Gallo, J. 2002, «Econométrie spatiale : l'autocorrélation spatiale dans les modèles de régression linéaire», *Économie & prévision*, vol. 155, n° 4, p. 139–157.

LeSage, J. et R. K. Pace. 2009, *Introduction to Spatial Econometrics*, Chapman and Hall/CRC.

Manski, C. F. 1993, «Identification of endogenous social effects : The reflection problem», *The Review of Economic Studies*, vol. 60, n° 3, p. 531–542.

Martinez, V. J. et E. Saar. 2001, *Statistics of the galaxy distribution*, CRC press.

Matheron, G. 1965, *Les variables régionalisées et leur estimation : une application de la théorie de fonctions aléatoires aux sciences de la nature*, vol. 4597, Masson et CIE.

Matula, D. W. et R. R. Sokal. 1980, «Properties of Gabriel graphs relevant to geographic variation research and the clustering of points in the plane», *Geographical analysis*, vol. 12, n° 3, p. 205–222.

Nadaraya, E. A. 1964, «On estimating regression», *Theory of Probability & Its Applications*, vol. 9, n° 1, p. 141–142.

Oppenshaw, S. et P. Taylor. 1979, «A million or so correlation coefficients», *Statistical Methods in the Spatial Sciences*.

Penttinen, A., D. Stoyan et H. M. Henttonen. 1992, «Marked point processes in forest statistics», *Forest Science*, vol. 38, n^{o} 4, p. 806–824.

Ripley, B. D. 1976, «The second-order analysis of stationary point processes», *Journal of Applied Probability*, vol. 13, n^{o} 2, p. 255–266.

Ripley, B. D. 1981, *Spatial Statistics*, John Wiley & Sons, New-York.

Statistique Canada. 2016a, «Fichier des limites», *Produit 92-160-X du Recensement*.

Statistique Canada. 2016b, «Guide de référence (deuxième édition) - Fichier des limites du Recensement», *Produit 92-160-G*.

Stevens Jr, D. L. et A. R. Olsen. 2004, «Spatially balanced sampling of natural resources», *Journal of the American Statistical Association*, vol. 99, n^{o} 465, p. 262–278.

Stoyan, D. et H. Stoyan. 1995, *Fractals, Random Shapes and Point Fields*, John Wiley & Sons, Chichester.

Strand, L. 1972, «A model for stand growth», dans *IUFRO Third Conference Advisory Group of Forest Statisticians*, vol. 72, Institut National de la Recherche Agronomique Paris, p. 3.

Strauss, D. J. 1975, «A model for clustering», *Biometrika*, vol. 62, n^{o} 2, p. 467–475.

Tiefelsdorf, M., D. A. Griffith et B. Boots. 1999, «A variance-stabilizing coding scheme for spatial link matrices», *Environment and Planning A*, vol. 31, n^{o} 1, p. 165–180.

Tobler, W. R. 1970, «A computer movie simulating urban growth in the Detroit region», *Economic geography*, vol. 46, p. 234–240.

Toussaint, G. T. 1980, «The relative neighbourhood graph of a finite planar set», *Pattern Recognition*, vol. 12, n^{o} 4, p. 261–268.

Upton, G., B. Fingleton et collab.. 1985, *Spatial Data Analysis by Example. Volume 1 : Point Pattern and Quantitative Data*, John Wiley & Sons.

Annexe A – Analyse descriptive des variables

Variable	Min	1er Q	Med	Moy	3ème Q	Max
Données de la Croix-Rouge canadienne						
Montant reçu ($ CAD)	0	600	740	1 566	1 744	33 506
NB in Casefile	1	1	2	2,41	3	9
NB 0 à 5 ans	0	0	0	0,11	0	3
NB 5 à 18 ans	0	0	0	0,40	0	6
NB 18 à 65 ans	0	1	2	1,54	2	7
NB + de 65 ans	0	0	0	0,35	1	6
NB sexe féminin	0	1	1	1,21	1	7
NB sexe masculin	0	1	1	1,20	2	6
Données externes						
Population 2016	259	516	687	861	894	5 210
Âge moyen	31,2	39,9	43,7	43,0	44,7	69,2
Taille moy ménages	1,5	2,3	2,3	2,4	2,5	3,2
Revenu med ménages	24 448	61 653	65 472	67 956	74 112	173 568
Taux d'activité	32,5	60,7	64,2	65,2	70,6	85,3
Taux d'emploi	29,4	56,1	57,1	60,6	65,9	78,5
Taux de chômage	0	5,6	6,9	7,2	8,5	20,0
Valeur med logements	174 252	200 348	219 911	247386	275 569	898 905
	Fréquences (sur 2406 foyers)					
Sévérité	Mineure	Modérée	Importante		Extrême	
	39	186	377		1 804	
	(1,6%)	(7,7%)	(15,7%)		(75,0%)	

TABLE 1 – Statistiques descriptives des variables disponibles pour le jeu de données *Montréal* de 2406 observations après pré-traitement.

Variable	Min	1er Q	Med	Moy	3ème Q	Max
Données de la Croix-Rouge canadienne						
Montant reçu ($ CAD)	75	600	880	1 655	1 800	33 506
NB in Casefile	1	2	2	2,65	4	9
NB 0 à 5 ans	0	0	0	0,15	0	3
NB 5 à 18 ans	0	0	0	0,48	1	6
NB 18 à 65 ans	0	1	2	1,69	2	7
NB + de 65 ans	0	0	0	0,32	0	6
NB sexe féminin	0	1	1	1,35	2	7
NB sexe masculin	0	1	1	1,30	2	5
Données externes						
Population 2016	362	516	698	967	1351	5 210
Âge moyen	31,2	39,0	43,4	41,2	43,7	43,7
Taille moy ménages	2,3	2,3	2,3	2,5	2,6	3,0
Revenu med ménages	60 800	61 952	66 133	70 536	79 616	108 864
Taux d'activité	58,1	60,7	64,6	66,3	73,8	81,70
Taux d'emploi	54,8	56,2	56,5	61,6	69,4	78,5
Taux de chômage	4	6,7	6,9	7,3	8,2	11,3
Valeur med logements	190 246	199 779	209 817	222 673	245 042	349 461
	Fréquences (sur 1 222 foyers)					
Sévérité	Mineure	Modérée	Importante		Extrême	
	0	89	0		1 133	
	(0%)	(7,2%)	(0%)		(92.7%)	

TABLE 2 – Statistiques descriptives des variables disponibles pour le jeu de données *Marthe* de 1222 observations après pré-traitement.

Tranche d'âge	0	1	2	3	4	5	6	7
0 à 5 ans	4326	263	108	8	x	x	x	x
5 à 18 ans	3804	433	340	88	32	6	2	x
18 à 65 ans	883	1697	1665	327	106	17	7	3
+ de 65 ans	3380	896	419	8	1	x	1	x

TABLE 3 – Fréquence des variables représentant les tranches d'âge sur le jeu de données total (Québec) de 4705 observations après pré-traitement. *Exemple : sur les 4705 foyers, 263 ont déclaré un enfant âgé de 0 à 5 ans, 108 foyers en ont déclaré deux et 8 en ont déclaré 3.*

Annexe B – Régression géographiquement pondérée

		Modèle 1 : variables CRC + sévérité		Modèle 2 : CRC + sévérité + StatCan	
Régression linéaire classique		R^2 adj = 0,046 AICc = 43 892		R^2 adj = 0,064 AICc = 43 903	
Noyau		Bande passante	Performance	Bande passante	Performance
Gaussien	Fixe	38 773	R^2 adj = 0,45 AICc= 43 893	66 399	R^2 adj = 0,063 AICc =43 854
	Adaptatif	2 172	R^2 adj = 0,046 AICc = 43 892	2 282	R^2 adj =0,063 AICc =43 854
Exponentiel	Fixe	1 679	R^2 adj = 0,15 AICc = 43 740	2 160	R^2 adj =0,16 AICc =43 775
	Adaptatif	1 650	R^2 adj = 0,048 AICc = 43 888	1 888	R^2 adj = 0,066 AICc= 43 851
Bicarré	Fixe	66 351	R^2 adj = 0,046 AICc = 43 892	66 427	R^2 adj = 0,064 AICc =43 855
	Adaptatif	1 688	R^2 adj = 0,054 AICc = 43 887	1 959	R^2 adj = 0,069 AICc = 43 858
Tri-cube	Fixe	60 441	R^2 adj = 0,046 AICc =43 892	66 427	R^2 adj = 0,063 AICc = 43 855
	Adaptatif	1 959	R^2 adj = 0,048 AICc = 43 894	1 984	R^2 adj = 0,067 AICc =43 863
Boxcar	Fixe	36 688	R^2 adj = 0,05 AICc =43 889	39 684	R^2 adj = 0,065 AIC =43 852
	Adaptatif	1 493	R^2 adj = 0,05 AICc = 43 895	2 190	R^2 adj = 0,064 AICc = 43 856

TABLE 4 – Performances des modèles de régression linéaire et de RGP selon les valeurs des bandes passantes optimisées pour le jeu de données *Montréal* avec la fonction `bw.gwr()` de *GWmodel*. Lorsque l'option `adaptive = TRUE`, la valeur de la bande passante est exprimée en nombre de points voisins et non en terme de distance.

Annexe C - Analyse surfacique

Définition du voisinage	Type de normalisation	H_0	I Global ~	valeur-p
Contiguïté Queen	En ligne (W)	rejetée	0,38	1,0e-07
	Globale (C)	rejetée	0,46	3,2e-13
	Uniforme (U)	rejetée	0,46	3,2e-13
	S. Varia. (S)	rejetée	0,43	4,8e-11
	Aucune (B)	rejetée	0,46	3,2e-13
Triang. de Delaunay	En ligne (W)	rejetée	0,38	2,2e-16
	Globale (C)	rejetée	0,40	2,2e-16
	Uniforme (U)	rejetée	0,40	2,2e-16
	S. Varia. (S)	rejetée	0,39	2,2e-16
	Aucune (B)	rejetée	0,40	2,2e-16
Sphère d'influence	En ligne (W)	rejetée	0,38	1,2e-08
	Globale (C)	rejetée	0,36	5,2e-09
	Uniforme (U)	rejetée	0,36	5,2e-09
	S. Varia. (S)	rejetée	0,37	4,3e-09
	Aucune (B)	rejetée	0,36	5,2e-09
Graphe de Gabriel	En ligne (W)	rejetée	0,28	1,7e-05
	Globale (C)	rejetée	0,27	2,4e-06
	Uniforme (U)	rejetée	0,27	2,4e-06
	S. Varia. (S)	rejetée	0,28	3,6e-06
	Aucune (B)	rejetée	0,27	2,4e-06
Voisins relatifs	En ligne (W)	rejetée	0,22	0,0015
	Globale (C)	rejetée	0,25	0,0002
	Uniforme (U)	rejetée	0,25	0,0002
	S. Varia. (S)	rejetée	0,23	0,0005
	Aucune (B)	rejetée	0,25	0,0002
Voisin le plus proche	X	rejetée	0,29	0,001
2 V. les plus proches	X	rejetée	0,34	1,0e-07
3 V. les plus proches	X	rejetée	0,33	4,8e-10
V. à la distance min	En ligne (W)	rejetée	0,23	2,2e-16
	Globale (C)	rejetée	0,15	2,2e-16
	Uniforme (U)	rejetée	0,15	2,2e-16
	S. Varia. (S)	rejetée	0,18	2,2e-16
	Aucune (B)	rejetée	0,15	2,2e-16

TABLE 5 – Indices de Moran (Global) selon la définition du voisinage et le type de normalisation. *Calculés avec la fonction* `moran.test()` *du paquetage spdep, sous l'hypothèse de randomnisation.*